Creativity in Comparative Analysis between Artists and Text-to-Images

Dario Tortorici
DISI, University of Trento, 248443
Trento, Italy
Email: dario.tortorici@studenti.unitn.it

September 16, 2024

***Abstract*—The topic of creativity has been of interest to philosophers, psychologists, and neuroscientists, as it is a fundamental human characteristic that facilitates problem-solving and drives artistic expression. This discussion has been further expanded with the growing interest in artificial intelligence, particularly in the field of generative AI. This paper presents a comparative analysis of the creative processes of human artists and text-to-image models, with a specific focus on the cognitive mechanisms and neural substrates that underpin the process. The research examines the cognitive aspects of creativity, including memory and the creative process, in order to compare human creative endeavours with state-of-the-art AI architectures. The objective is to identify the similarities and differences between the two systems and to propose potential implementations within existing architectures, with a view to improve their creativity. The study emphasises the collaborative nature of creative expression and highlights the symbiotic relationship between human input and machine output in the realm of text-to-image generation. This is done to contribute to the ongoing dialogue about the nature of creativity and its manifestation across modalities.**

I. Introduction

Since the 1950s and 1960s, computer scientists have been developing computer programs to simulate human cognitive processes. However, defining and measuring creativity has been a persistent challenge, leading to ongoing debates that highlight the inherent complexity of the phenomenon. Creativity has long been a construct of interest to philosophers, psychologists and neuroscientists because of its mysterious nature and recognition of a crucial aspect of human endeavor, as it promotes innovation, adaptability, and problem solving in various domains [1], [2]. More recently, the emergence of artificial intelligence (AI) has added a new dimension to the study of creativity, as AI systems, in particular deep neural networks, draw inspiration from the human brain in their architecture and functioning. Although AI has made remarkable progress in areas such as image generation and natural language processing, replicating human creativity remains a challenge. This may be due to the inherently elusive nature of creativity, which defies easy definition and quantification. Another problem is that current evaluation methods do not take creativity into account. Furthermore, the assessment of creativity is now carried out by humans and may be susceptible to the introduction of potential biases due to the characteristics of the creator. It is often assumed that generative AI puts less effort into creating a given artefact than humans, resulting in less creativity being attributed to generative AI producers [3], [4].

This review examines the cognitive processes and neural correlates of human visual creativity and their contribution to the creation of visually compelling content. Furthermore, the paper examines AI-driven creativity and their evaluation, specifically when AI generates visual content based on human textual input. The objective of this paper is to clarify our comprehension of human cogni-

tion and artificial intelligence by examining the similarities and differences in creativity between humans and AI. Furthermore, suggestions will be put forth regarding the incorporation of creativity as a modulating factor into existing AI architectures.

Defining Creativity: The assessment of creativity remains a contentious subject in the field of creativity research. Over time, numerous frameworks and assessment tools have been devised with the aim of establishing standardised definitions and metrics for the topic. These frameworks aim to provide a meaning and tools for evaluating creativity, facilitating comparisons not only among individuals but also between different modalities of creative expression, including generative models [5] and biological brains [6], [7]. In the field of research, creativity is often defined as the production of something both novel and useful within a given social context [8]. This definition serves to illustrate a consensus across a range of disciplines. Nevertheless, despite this consensus, the execution of scientific tests to measure creativity remains challenging due to the complexity of adhering to established standards. A significant number of the tests reviewed in this study rely on evaluations made by human judges, who inevitably bring their own subjective interpretations of significance [9]–[12]. Psychometricians emphasise that for these tests to be meaningful, they must meet the criteria of validity and reliability [13]. Validity ensures that a test accurately measures the intended aspect of creativity, whereas reliability ensures consistent results. The intrinsic subjectivity of evaluating creative output highlights the necessity for at least one creativity test that can offer clear avenues for performance enhancement, in accordance with these principles. Consequently, in a subsequent sections of this paper, we will also discuss some possible metrics to implement a more structural approach to mitigate the problem of this subjectivity.

Refining Comparative Parameters: To conduct a thorough comparative analysis between these two systems, it is essential to define specific parameters to ensure the accuracy and validity of the results. This comparative examination is focused on a specific creative task, rather than a broad examination of creativity as a whole. The approach is based on the premise that human creativity is contingent upon the specific task at hand [14]. Therefore, it is necessary to make specific comparisons between each type of artificial architecture and the biological one for each creative activity. This investigation focuses on the increasing use of AI systems that can transform text into visual representations. To ensure a fair and informative comparison, it is important to consider the unique characteristics, limitations, and capabilities of each system. This is essential to avoid any appearance of bias or superficiality in the analysis. The study aims also to investigate also whether artificial neural networks can exhibit creative agency similar to that observed in human cognition.

II. Human architecture

Creative Process: The most basic model of the creative process is a two-stage model, which is sometimes referred to as the baloon model. This model involves an expanding stage of divergent thinking, where a multitude of possibilities are generated, followed by a stage of convergent thinking, where the focus is on identifying the one best idea [15]. A comparison can be drawn between the two-stage model of the creative process and the functioning of text-to-image models. In these models, the model generation of potential visual interpretations based on a specified text prompt may be considered analogous to divergent thinking. The refinement stage is analogous to convergent thinking, whereby the optimal image that most accurately represents the text input is produced from the entire range of possibilities. The creativity process has been the subject of multiple studies, which have demonstrated that creativity is not a two-stage models but tends to occur in a sequence of more phases [16]–[20]. To standardise these researches, we can say that a general framework about the creative process in the human brain is composed by eight distinct stages [13]. Firstly, the individual identifies and frames the problem. This first step involves the task of identifying a problem that is not only significant, but also framed in a way that is conducive to a possible creative solutions.

Subsequently, individuals employ the acquisition of relevant knowledge to the identified problem. Creativity relies heavily on the accumulation of expertise, mastery, and practice relevant to the given domain [21]–[23]. Once individuals have acquired the requisite knowledge, they begin to gather a variety of information that may be pertinent to the topic at hand. This differs from the previous step, as it involves information that is not strictly related to the topic. The incubation phase then begins to play a role. During this phase, the unconscious mind processes and associates the acquired knowledge and seemingly unrelated information in unpredictable ways, setting the stage for creative insights to emerge. After the accumulation and incubation, individuals generate a multitude of ideas, which we have defined as divergent thinking [15]. During the creative process, individuals must exercise discernment by selecting the most promising ideas. Given the plethora of potential solutions, successful creators must have the ability to apply relevant criteria to identify ideas worthy of further pursuit, which is known as convergent thinking. Finally, creativity is externalised using various materials and representations. In the artificial networks, the stages of information gathering and incubation are currently absent from the process in artificial networks. The implementation of these processes may facilitate the regulation of the creative process and, subsequently, the creativity of the output. As you notice, the conscious mind plays an important role in the creative process and therefore is difficult to compare it with an artificial model. The initial three stages are predominantly conscious and directed. Even the fourth, incubation, only occurs in the context of ongoing conscious work. This explains why incubation can only be beneficial if one has previously invested significant effort into a problem and then continues to work on it afterwards [24]. As our primary objective is to examin commissioned works of art, the first point is somewhat sweetened. The artist is presented with an already established problem and is therefore only able to contribute to its formulation. However, it is worth noting that creativity scholars have found that exceptional creativity often occurs when individuals work in areas where problems are not predefined, this is because these scenarios allow for greater freedom and demand more innovative thinking. The ability to formulate effective questions is a key factor for success in these areas [25], [26]. This implies that a commission may be less creatively stimulating for an artist than initiating a new project independently, which is a reasonable assumption, because the artist lacks of first intention to expression. In fact, research identified twelve attributes that individuals consider when evaluating ideas for expression [27]. These attributes includes the perceived risk level of the idea, its ease of understanding, originality, thoroughness (whether it provides detailed implementation steps), complexity, alignment with existing social norms, likelihood of success, ease of implementation, potential benefits to a broad audience, alignment with desired societal goals, as well as the time, effort, and complexity required for implementation. In our analysis the process, which is established within a committee, is agreed with the contractor explicitly or implicitly through feedback on the final work. Implementing these attributes as hyperparameters can be the key to allowing the end user to choose how much creativity they want in their output.

Role of the memory: The Gestaltist theory suggests that the mind is capable of sudden restructuring, leading to a moment of insight known as the Eureka moment [28]. However, studies have disproved this theory, showing that the mind gradually approaches the correct solution [29]. This is consistent with the associationist theory, which postulates that creativity arises from the convergence of pre-existing ideas. This theory elucidates the manner in which these associations facilitate the formation of novel connections and the generation of innovative ideas. This theory posits that creative ideation is initiated by the functioning of semantic memory, where concepts, ideas, and experiences are interconnected to form a network of associations. It is postulated that knowledge is organised in a structured manner, with concepts being related to one another. Search processes operate across this semantic space, resulting in memory search, retrieval, and creative combination. The theory of creativity

suggests that individuals with higher creativity possess a more extensive semantic memory structure, enabling them to conduct a broader search within their memory [21]. Humans have different types of memory and either short-term memory (STM) and Long-term memory (LTM) are involved. In particular: semantic memory [30], episodic memory [31], association [32] and combination [33] have been identified as cognitive components of the creative process. STM is involved in creative processes because they require the temporary storage of information [34], [35]. LTM has been linked to creativity because it stores information about prior knowledge [36]. LTM can be classified into two types: declarative memory and non-declarative memory. Declarative memory is the type of memory that can be consciously accessed, the most related to the activity and is further divided into semantic memory and episodic memory. A review of the literature on semantic memory search processes reveals a correlation between clustering, defined as the grouping of related concepts during memory retrieval, and divergent thinking. Furthermore, the capacity to transition between disparate categories or domains of thought is associated with the integration of remote associates. These processes are of great importance for the development of higher levels of creative thinking. The findings indicate that clustering is more strongly associated with divergent thinking, whereas switching is more closely related to convergent thinking and the efficiency of the semantic network [37]. Another study suggests that the semantic memory network of people with low creative ability appears to be more rigid than that of people with high creative ability, in the sense that it is more spread out and divided into more subparts [38]. Conceptual combination is the mental act in which imagination brings concepts together to produce new ideas in creative processes [39]. These creative combinations probably have properties that aren't held by the component concepts. This retrieval and combination of previous memory processes can stimulate imagination [31]. Other findings have also been promoted, such as that highly creative people are more likely to use remote association during a creative process [40]. Memory is one of the fundamental elements of creativity [41], [42]. Creativity cannot occur ex nihilo, but is a process in which novel ideas are generated by searching [43], interacting [44] and associating [45] existing memories. Objects, rather than features, are generally considered to be the elementary building blocks of our visual representations, not only for perception [46], [47] but also for visual working memory (VWM) [48], [49]. Despite its importance, the majority of contemporary architectural designs do not incorporate an explicit network that emulates the human memory process. The introduction of such a network may facilitate the introduction of creativity as a parameter. The Hopfield Neural network (HNN) [50] is know to his capacity to emulate human memory assotiation process. A study introduces the use of them as a tool to emulate the creative process through concept association [51]. This approach is based on the architecture of neural networks, which mirrors the way human memory works, where multiple neural units are activated simultaneously in response to given stimuli. Using modern HNNs, it was possible to simulate in a discrete and asynchronous way the ability of human creative thinking to make meaningful connections between seemingly unrelated concepts. The research demonstrated success in implementing a neurocomputational framework for creativity-based semantic associations using both binary and modern HNNs [52]. However, the study is limited by the low memory and nodes ratio, as only approximately 138 vectors can be retrieved from storage for every 1000 nodes [53].

Brain flow and structure: Creativity is not specifically associated with any single area of the brain [54]. Thus, understanding the distribution of information throughout the brain is considered a crucial factor [11]. As the source of new ideas, memory has been identified with the activity of amygdale by functional magnetic resonance imaging (fMRI) [55]. Results from numerous fMRI studies also consistently highlight the central role of the prefrontal cortex in both hemispheres [56]–[60], presumably due to its involvement in working memory and executive attention processes. This observations aligns with the framework of asso-

ciative thinking, wherein semantic understanding of words precedes the activation of memory for visualisation. Recent research has called into question the prevailing view of the lateralisation of brain activity during the creative process. New evidence suggests that the frontal lobe, which is responsible for executive functions such as planning and action control, is involved in creativity. This challenges the prevailing notion that creativity is predominantly associated with the right hemisphere. To disproof the hypothesis, fMRI was used by researchers to evaluate neural activity in individuals during a visuospatial creativity task [61]. The task is known to rely on divergent thinking, which is a hallmark function typically associated with the right hemisphere. A further study corroborates these findings by comparing electroencephalogram data with existing research on brain activation during creative cognitive tasks. It was found that remote association is associated with frontal lobe activity [62]. This contrasts with some previous studies that associate it with temporal lobe activity [63]. The variation in outcomes could be attributed to the engagement of semantic and episodic memory, indicating that distinct brain regions may be activated during remote association, potentially influenced by the type of induction task and the method of stimulus presentation. Also brain wave activity are involved, exprecially alpha [64], theta [65], and gamma waves [66]. In addition, leftward gaze shift when participants were required to think about an original idea and pupil dilatation [9]. Furthermore, when discussing the specific process of drawing, additional brain structures come into play that were not previously studied in the context of divergent thinking. These include motor and somatosensory areas, as well as cortical regions involved in spatial and auditory perception [67]–[69]. It is important to highlight that the aforementioned insights are not reflected in the current state-of-the-art methodologies. This is due to the current lack of sufficient knowledge to implement artificial counterparts. Moreover, it is uncertain to what extent this has an impact on the final result.

Divergent thinking is commonly measured using the Alternative Uses Test [15], which evaluates a person's ability to generate multiple uses for a common object or to think of novel and unusual uses for everyday items. The test in question is unable to fully represent the artist's creative process, as it does not involve physical movements or imagery. Furthermore, tasks such as imagining an apple, performing a mental rotation, or engaging in a sporting activity are often referred to as 'imagery', despite being very different. Studies have demonstrated that the visual imagery employed by artists when contemplating subjects is analogous to perception, albeit with a diminished level of activity. This imagery is also influenced by the strength and duration of the stimulus. This can be attributed to their similarity in the brain processes. In fact, there is a degree of overlap between the brain regions involved in imagery and those involved in perception of previous experiences [70]. Another research has shown that participation in visual arts education can lead to changes in brain structure, particularly in the density of grey matter [12]. Moreover, in cognitive drawing, operationalised by internally cued drawing stimuli or objective drawing content, there is evidence to suggest that this process is associated with the activation of the prefrontal and cingulate cortices [71].

The AI counterpart is based on the premise that the mind can be viewed as a computational device. This approach limits the study of the human mind to cognition, excluding emotions, motivations, and irrationality, which are also involved in the creative process. Studies reveal that mood states significantly influence creativity. Positive-activating moods, such as happiness, and negative moods, such as sadness and depression, have a positive impact on the process. Conversely, relaxation, anger, and anxiety have a negative influence [72]–[74]. In addition, machine learning (ML) algorithms rely on their ability to represent numbers with a high degree of resolution and accuracy. This is difficult or impossible with biological neurons. Moreover, the more accuracy needed, the slower a neuron-based system will run [75]. It is impossible that any biological brain could implement the numerical precision required by ML in a sufficiently rapid manner to be useful. From our perspective, the

production of a ML model is not creative due to the mathematical precision involved. While we also perform mathematical calculations, they are approximate. In contrast, machines aim to achieve greater precision, which is not aligned with our experiences of creativity. Neurons operate at a much slower pace than electronic signals, firing at a maximum rate of approximately 250 Hz. This structural difference serves to highlight the dissimilar working processes. The maximum number of values that can be represented by one neuron firing is between 10 and 100 unique values. ML algorithms require a greater degree of precision than this, as the underlying concept of gradient descent assumes the existence of a continuous gradient surface. In addition, dendrites have the ability to perform XOR operations [76], which perceptrons [77] cannot. Therefore, numerical comparisons between the two respective neurons are not valid. Indeed, research has demonstrated that in order to accurately replicate the behaviour of a human neuron, it is necessary to implement approximately five to seven CNN layers [78]. This implies that a single neuron in our body can be compared to an entire network.

III. Artificial Neural networks

Prior to discussing the various types of text-to-image architectures, it is essential to delineate the distinction between these models and the processes involved in training and learning how to utilise them in relation to human beings. In the context of the training process, the artificial neural network is able to emulate the philosophy of deliberate practice [23] and the incorporation of feedback, as exemplified by the backpropagation mechanism. However, the absence of a tangible layer structure in a biological system precludes the existence of a mechanism that could be employed to dictate the weight of any specific synapse. Consequently, in order to reduce the weight of a synapse in our brain, the target must fire shortly before the source. In order to effect any meaningful change in a synapse weight, a number of repetitions must be performed. Additionally, the connections in the brain are not organised in the orderly layers as artificial neural networks [79], which requires some modification to the basic perceptron algorithm. This may prevent backpropagation from working at all. Furthermore, during the execution phase, it can be reasonably assumed that the quality of the prompt entered into these models also influences their ability to produce images. This is arguably true even for humans. However, for artificial networks, the concept of creativity is more bounded to the prompt, as they lack a particular style of initiative to modify or intepret the initial commission.
In the field of text-to-image synthesis, the three main methodologies - Generative Adversarial Networks (GANs), autoregressive methods, and diffusion models.

Generative Adversarial Networks

GANs [80] uses two neural networks that work together: one generates artificial images using a random noise vector, while the other determines whether the input image is real or artificial by comparing it to samples from the training data. GANs are renown for their single-step formulation and efficiency. However, despite efforts to scale them up for handling large datasets [81], diffusion models have shown significant results that surpass the quality of GAN models [82]. This approach emulation of previous work can be seen as good simulation of the process of an artist learning by creating an image and then receiving feedback. However, there is no correlation between the human and artificial structures involved.

Autoregressive

Autoregressive methods for text-to-image synthesis are able to capture the details and global coherence by modeling the conditional probability distribution of image pixels based on textual descriptions. The models linearize 2D images into 1D sequences of patch representations using a Transformer Model [83]. This sequential generation differs markedly from the parallel approach employed by GANs, as autoregressive models generate images pixel by pixel. Although this sequential process results in superior global image coherence, it is more computationally expensive during both training and inference compared to GANs. Analogies

with humans are the ability to focus on both the global composition and details of the image.

Parti: Parti [84] is a Google research project that utilizes standard Transformers for all of its components, which are the encoder, decoder, and image tokenizer. The model is composed of two stages: an image tokenizer and an autoregressive model. In the first stage, the tokenizer is trained to convert images into a sequence of discrete visual tokens for training and reconstruction purposes. The second stage trains an autoregressive sequence-to-sequence model that generates image tokens from text tokens. This process is analogous to the manner in which the human brain processes language. Initially, the brain deconstructs spoken or written words into fundamental phonetic or semantic units. Subsequently, it reconstructs meaning [85]. In this type of model, an increase in prompt complexity may result in the occurrence of errors, including colour bleeding, omission, hallucination, duplication of details, displaced positioning or interactions. To obviate these errors, it is essential to augment the number of tokens in a proportionate manner in order to generate a well-structured output. Furthermore, the Parti autoregressive solution incorporates the semantic content of an image into the encoder process. The model employs cross-attention, a mechanism that enables the model to focus on pertinent information from the encoder while processing another sequence for the decoder. In particular, the text encoder embeddings are employed as conditioning for the image decoder, which predicts one image token after another. This approach facilitates the handling of image coherence and long prompts, provided that the number of tokens is adjusted in accordance with the aforementioned discussion. This cross-attention mechanism has strong analogies with the human ability to process relations and reason in terms of analogies [86].

CM3Leon: Meta has developed CM3Leon [87], with the objective of balancing the inference time with global image coherence. This is achieved through the utilisation of retrieval-augmented pretraining on a comprehensive, heterogenous multimodal dataset. The CM3Leon model achieves state-of-the-art performance in text-to-image generation using five times less training compute than comparable methods. In contrast to Parti, CM3Leon employs a decoder-only transformer architecture that does not incorporate an encoder. In order to merge the two token vocabularies in the decoder, a break token is employed to indicate the point at which the text tokens cease and the image tokens commence. The CM3Leon model is capable of accepting multi-modal inputs, thereby making it a versatile model that is suitable for infilling and autoregressive generation tasks for both images and text. This approach facilitates not only text-to-image conversion but also image-to-text conversion. This approach is analogous to the human capacity to convert one modality to another. Indeed, the transformer decoder is capable of accepting either text or image as input, with the vocabulary being merged. The mask system allows the user to specify which modality is to be input and which is to be output. Although the CM3Leon architecture exhibits a capacity for multimodal switching comparable to that observed in the human brain, the latter does not possess a fully reversible process between these modalities.

Diffusion Models architectures

Diffusion models [88], [89] have become popular in image generation due to their strong performance and relatively low computational cost [90]. The main concept, which draws inspiration from non-equilibrium statistical physics, involves a gradual and systematic breakdown of structure in a data distribution through an iterative forward diffusion process of Gaussian noise. Subsequently, we acquire knowledge of a reverse diffusion process that reinstates structure in the data. The architecture of the denoising network can vary, but many diffusion models use a U-Net architecture [91]. At the time of inference, the reversal of the forward paths of data towards noise enables the generation of data from noise. Diffusion models are therefore said to emulate the human process of sketching and refining a result. In particular, the noise predictor estimates the noise present in the image, which is then subtracted from the image. This process is not conducted in a single instance, as this was demonstrated to be less

effective; instead, the process is repeated on several occasions, with a gradual reduction in noise until a clean image is obtained. This denoising process is referred to as sampling, as diffusion models generate a new sample image at each step. The sampling method, known as the sampler, regulates the level of noise at each sampling step. The highest noise occurs at the first step and gradually decreases to zero at the final step. The objective of the sampler at each step is to generate an image with a noise level that corresponds to the noise schedule. This is for human a subjective process and even for artificial neural networks, with multiple noise samplers existing, the optimal one depends on the task [92]. For istance, according to some study, a linear schedule is not well-suited for low-resolution images [93]. The samplers labelled 'Karras' use the noise schedule recommended in the Karras article[94]. The study suggested that the noise step sizes should be smaller towards the end compared to the standard. This is argued to improve the quality of images. Also, for the purpose of reproducibility, it is important for the image to converge, even though not all samplers may converge and can run indefinitely. To generate minor variations in images, a different variational seed is used, which refers to the initial noise. Furthermore, it has been observed that there are differences in the application of noise, whether in pixel or latent space. Instead of adding noise directly to the pixel space of images and then denoising them, a latent diffusion model takes a different approach. It encodes the images into a latent space where the noise is applied and denoised. After this process, the image is decoded from the latent space, yielding an image that closely resembles the original. This method is referred to as a latent diffusion model [95], and it offers a significant improvement over the basic diffusion model. One of the most significant advantages is the increased speed, as the process is many times faster than denoising the raw, uncompressed data directly. It should be noted that this latent space sampling approach is a computational-saving technique that is not feasible for humans. Moreover, certain models exhibit a fixed width and height production of the image, which contrasts with the capacity of humans to adapt their representation according to their imagination. Additionally, the two systems are distinct in terms of their underlying sampling philosophies. For humans, the generation of different outputs based on an identical prompt is not related on the initial noise, comparable to the initial canvas, despite the support's capacity to influence the choices and, consequently, the final result.

CLIP: Before diving into OpenAI's models, it's essential to first examine the role of "Contrastive Language-Image Pre-training" (CLIP) technology, which has significantly influenced advancements in the text-to-image domain. Although CLIP is not directly implemented in DALL-E 2 [96] or DALL-E 3 [97], it has nonetheless informed the foundational principles of these models and is a key component in other diffusion models. The distinguishing feature of CLIP is its capacity to comprehend the interrelationship between images and text. The model comprises both an image encoder and a text encoder, trained on a substantial dataset comprising 400 million images and their corresponding captions. The objective of CLIP is to generate comparable embeddings for matching images and captions, while producing disparate embeddings for non-matching pairs. This is the method employed by diffusion models to emulate the human capacity for processing relations. For the image processing, CLIP utilizes a convolutional neural network (CNN) as its vision encoder. The CNN extracts high-level features from images, which are then refined through additional layers. On the textual side, CLIP employs a transformer-based text encoder that captures and encodes semantic information in a format that aligns with the image representations. Both images and text are processed concurrently in CLIP, enabling the model to learn to associate image representations with their corresponding textual descriptions. However, whereas human knowledge is structured and concepts are related by proximity, reflecting associative theory, in CLIP's embedding, the intrinsic meaning associated with a word or concept is replaced by its correlation with visual representations. This approach is effective for capturing high-level semantic relationships, but it does

not explicitly model the internal structure of objects or their specific attributes. The model is capable of discerning similarities between concepts throughout the text embedding process; however, it lacks the capacity to explicitly encode the specific attributes that distinguish various breeds or detailed features [96]. This discrepancy highlights the distinctions between the human semantic memory and the CLIP embedding process, particularly with regard to the replication of attributes associated with subjects in the output.

DALL-E 2: OpenAI employed the CLIP architecture as a preliminary step in developing the DALL-E 2 architectures for text-to-image synthesis. Consequently, the model is also referred to as unCLIP, as it accepts input from the CLIP embedding and generates the final result. The unCLIP model is based on a prior and a decoder. The prior takes as input the CLIP image embedding and produces another image embedding. This process is the main difference between the first DALL-E [98] architecture and the second implementation. The prior process was added because it offers more diversity and correlation of text input. While this could be achieved through an autoregressive or diffusion process, the diffusion method was favoured for its efficiency. For the implemented prior, the researchers employed the cosine noise schedule [93]. The schedule addresses early noise issues in the forward noise process and improves sample quality and training efficiency as the number of timesteps increases. The decoder is a previous model, Glide [98], modification. The decoder model takes two inputs: the new latent vector generated by the prior, representing the target image, and the initial caption text. The caption text is incorporated into the diffusion process to refine the image generation and align it with the text's semantics. This integration increase the relevance of the generation. In particular, the DALL-E method appears to be the most natural emulation of the human brain among all the state-of-the-art architectures. The text-encoding process converts natural language into a graphical representation of it, which is aligned with the associative thinking process, as previously stated. Subsequently, the prior applies modifications to this vector in order to enhance diversity and, consequently, creativity. This aligns with the initial stage of the creativity process, namely problem framing, which has been demonstrated to improve the potential for creative outcomes. Indeed, the phenomenon has been observed in studies of creativity in art students. These studies have found that students produce more creative work when they take the time to modify the spatial representation of their objects [99].

DALL-E 3: DALL-E 3 [97] is an improvement over its predecessor because it includes an additional process to elevate captions accuracy. DALL-E 2 was trained using self-supervised images and their corresponding captions. However, the captions often lacked descriptive detail. The new process trains CLIP with another model that analyzes the image and creates a synthetic descriptive caption, improving the quality of the data fed. This addition improves the final performance, but has no effect on our discussion.

Stable Diffusion: Stable Diffusion [95], introduced by StabilityAI in 2022, was the first model capable of generating images based on text prompts or other conditioning inputs. It processes images using a variational autoencoder and diffusion in the latent space. The model has since been improved, with the third version incorporating Rectified Flow [100], that directly links data and noise using a stochastic differential equation for sampling. The architecture involves encoding a caption with multiple encoders, including two CLIP models and T5, to create an intermediate representation that is concatenated with the outputs. This model does not facilitate the generation of novel ideas or insights that parallel the functioning of the human brain. Its reliance on purely mathematical guidance to expedite the search for solutions is not conducive to the emergence of creative thinking.

SDXL: SDXL [101] is a StabilityAI model architecture that merges the image quality of diffusion models with the speed of GANs through a process of distillation [102]. The model employs Adversarial Diffusion Distillation (ADD), which distills pretrained diffusion models into high-fidelity outputs by reducing the multi-step sampling process to

just 1-4 steps, while maintaining high sampling fidelity and potentially improving overall performance. This method employs a large-scale image diffusion model as a teacher signal, with a student trained as a denoiser optimising two objectives: an adversarial loss, which aligns generated samples with real images; and a distillation loss, which matches the targets of a frozen diffusion model teacher. However, the model has limitations, including difficulty in generating pure black or white [103] and placing subjects on solid backgrounds [92]. If there were a loss in creativity scoring, this could be a method of creating a smaller model from a larger one, which would emphasise creativity without losing quality output standards. This could be a good case because for creative compositions the model drawbacks are irrelevant and we do not yet know whether the size of the model improves creativity.

Imagen: Imagen is the diffusion model developed by Google [90]. It is architecturally similar to Stable's latent diffusion model. Its main innovation is that it uses an extensive pre-trained NLP model called T5-XXL, which is already trained, instead of using a text encoder trained on image captions. This allows the model to understand language more deeply, as it has seen more diverse and complex texts than just image captions. Nevertheless, the model still has difficulties with feature blending, omission or duplication of details, displaced positioning of objects, counting and negation in text prompts, as the presence of the word is interpreted as a feature that has been requested and displayed. This gives us a clear indication of how different it is, despite the emulation attempt, to capture the real working system of our natural language processing.

MidJourney architecture: Regrettably, despite its significant role in the current state of the art, little is known about thi s architecture due to the absence of published papers.

Playground: Playground v2 [104] is a Latent Diffusion Model that employs two fixed, pre-trained text encoders (OpenCLIP-ViT/G and CLIP-ViT/L) and follows the same architectural framework as SDXL. The model addresses the challenge of accommodating a range of aspect ratios in image generation. The importance of preparing a balanced bucketed dataset is highlighted. Furthermore, the study examines the pivotal function of aligning the outputs of the model with human preferences. This guarantees that the generated images will align with human perceptual expectations. To this end, they developed a system that enables the automatic curation of a high-quality dataset from multiple sources based on user ratings on their platform. The model was subjected to an evaluation process, during which it was demonstrated to outperform SDXL in all aspect ratios and to exhibit superior aesthetic quality when compared to DALL-E, MidJourney 5.2 and SDXL. Furthermore, the same type of evaluation process may be conducted with regard to creativity.

Hybrid architectures

The objective of these models is to impose constraints on models based solely on text, as well as on architectures that accept additional information as input. This type of model shifts the focus of the discussion from the architectural design and creative output to output alignment with human constraints. The incorporation of supplementary data allows the majority of the creative process to be completed by the human designer. As previously discussed in the context of the creative process, this results in the output becoming primarily a matter of execution.

Meta make a scene method: This architecture [105] is designed to increase accuracy and relevance through the use of meta-feedback. This approach allows users to sketch their desired output and refine the generated images. Meta's method integrates several features into its classical autoregressive paradigm (CM3Leon) to address challenges such as object positioning and spatial interpretation of the described scene. First, Meta introduces a scene layout control mechanism, which complements textual input, improving structural consistency and quality while allowing scene editing. The use of a scene composed of semantic segmentation groups provides additional global context and conditioning cues during image generation. This is achieved by explicitly guiding the model to generate images that better match human preferences. Second,

the tokenisation process is refined by incorporating domain-specific knowledge about key image regions, such as faces and salient objects. This enhances the overall representation of the token space and improves the quality of generation and alignment with textual input.

Styledrop: The synthesis of image styles that utilise specific design patterns, textures, or materials is impeded by the challenges of natural language ambiguity and the presence of out-of-distribution effects. Styledrop [106] uses a transformer-based model for text-to-image generation, specifically leveraging Muse, [107] a transformer model capable of modelling discrete visual token sequences. This architectural approach offers a distinct advantage over diffusion models such as Imagen and Stable Diffusion, particularly in the context of learning fine-grained styles from single images. It involves taking a representation of the caption through the Muse model and fine-tuning it with a specific style, drawing from one or more images as a reference. Styledrop employs an iterative training framework to raise model performance over successive iterations.

IV. Creativity Measurements

As we've already discussed, we have difficulties in assessing creativity, from establishing a definition to subjective human judgement. As a result, optimizing creativity is difficult if it cannot be accurately measured.

Actual ANN Evaluation metrics

Current metrics in text-to-image architectures are inadequate for assessing creativity because they primarily measure the similarity between generated images and real ones. These metrics often reward systems for producing images that closely mimic real data, which can result from slight modifications to the input, leading to problems such as memorisation or mode collapse. This is of course not completely wrong, in evaluating the usefulness of a creative definition, it is implicit that the image in question should be meaningful. While hallucinations or random noise may possess novelty, they are devoid of usefulness. As a result, these metrics do not effectively capture the creative quality of the output and focus on replication rather than true innovation.

a) Inception Score: The Inception Score (IS) [108] is a widely used metric for evaluating the quality and diversity of images generated by AI models. The IS assesses image quality by checking how clearly identifiable and realistic the generated images are, and it measures diversity by assessing the variety of different images produced. This process is not aligned with our definition of quality, which is to resemble realistic representation, and limits some possible artistic representations, and diversity is not really what we mean by novel, but a way of representing the entire training set without mode collapse. Despite these drawbacks, the IS remains popular due to its simplicity and computational efficiency.

Fréchet Inception Distance: The Fréchet Inception Distance (FID) is a metric that employs a combination of the Fréchet distance [109] and the Inception score. The metric compares the distribution of generated images with that of a set of real images. Consequently, the FID best score is indicative of a greater degree of sample variety within the set. In contrast, the best IS score tends to demonstrate superior image quality within individual images. As with the Inception score, this metric assesses the generated images in terms of their visual fidelity and diversity in comparison to real images, rather than in terms of creativity. Despite its extensive utilisation within the industry, it has been demonstrated that FID score computed for a finite sample set is not the true value of the score [110] and may diverge from the assessments of human raters [111].

Contrastive Maximum Mean Discrepancy: In response to the limitations of FID, Google researchers have proposed an alternative metric known as the Contrastive Maximum Mean Discrepancy (CMMD) [111]. This metric is based on the utilisation of CLIP embeddings, in conjunction with the maximum mean discrepancy distance. This estimator is unbiased and does not make any assumptions regarding the probability distribution of

the embeddings in contrast to FID. Furthermore, it is sample-efficient. In comparison to FID and IS, it offers a more accurate and nuanced assessment of image quality and diversity. However, still requires complementation with human evaluations.

Text-image scoring methods: Text-image scoring methods are capable of evaluating the degree of correspondence between a generated image and a text prompt. For example, CLIP itself or Bootstrapping Language-Image Pre-training (BLIP) [112] may be cited. BLIP employs a multimodal encoder-decoder architecture, comprising a unimodal encoder for the reconciliation of visual and linguistic representations, and an image-based text encoder and decoder with a variety of attentional layers for the meticulous assessment of compliance with the prompt. This category of metrics is also concerned with the extent to which the generated output adheres to the prompt. Nevertheless, it is not always the case that these methods align with human preferences and perceptions.

Learned Perceptual Image Patch Similarity: Learned Perceptual Image Patch Similarity (LPIPS) [113] has been developed with the objective of more accurately gauging the high-level semantic similarity between images. The SqueezeNet [114] architecture is employed to extract deep features from images, which are then compared to ascertain perceptual similarity. Despite being trained on human judgments of perceptual similarity, LPIPS is susceptible to adversarial attacks that can deceive its neural network and consequently affect its final judgment.

ImageReward: The ImageReward model [115] is a text-to-image human preference reward model that integrates human preferences throughout the diffusion process. In comparison to CLIP and BLIP previously discussed metrics, ImageReward demonstrates superior performance in human evaluation. One of the most significant advances in ImageReward is the incorporation of Reward Feedback Learning (ReFL), which facilitates the calibration of diffusion models to closely align with human preferences. Following the generation of images, human evaluators assess the results based on four parameters: overall satisfaction, adherence to the prompt, aesthetic quality, and meta-feedback. This is especially beneficial during the final denoising stages of the diffusion process, where direct feedback learning can be employed even in models that do not inherently provide probability estimates for their outputs. The ImageReward model thus simulates the process of commissioning an artwork, where the input of human preference feedback is of crucial importance in the refinement and enhancement of the output of text-to-image models. Furthermore, this is the first measure that considers the creativity required by the prompter. This approach, however, necessitates that the user provide guidance at each stage of the process, ensuring that the output aligns with their expectations.

Possible metrics for the creativity evaluation

The following section is based on conjecture and is not designed to replace the established metrics; rather, it is intended to complement them, providing a means of gauging the degree of creativity inherent in the architectural design and potentially facilitating the modularisation of creative output. It is not the intention of this proposal to supplant the established metrics for evaluating image quality, adherence to the prompt, and fidelity of the training set. However, the proposed metrics may assist in overcoming the inherent limitations of AI models, which are designed to generate outputs that reflect the data with which they have been trained [116].

Conceptual Blending Score: A conceptual blending score can be used to assess the capacity of a model to integrate disparate concepts in a unified manner, which is frequently regarded as a hallmark of creative cognition. This may be quantified by identifying the concepts and evaluating the semantic consistency of the combined concepts in relation to the text prompt and their image representation. The metric could analyse the extent to which the model integrates these concepts without compromising their intrinsic characteristics.

Divergent Thinking Score: This metric would assess the diversity and potential divergence of concepts in the generated images in comparison to the prompt. It would not, however, take into account any concepts that were not expressed in the

prompt. The conceptual distance may be quantified by a knowledge graph constructed to represent the relationships between the concepts in the training set.

V. Discussion

What constitutes a creative AI product, and is this creativity required to mirror the characteristics of human creativity? Although AI systems are able to produce images that exceed the technical constraints of human creation, it is pertinent to question whether AI-driven creativity must adhere to the standards of creativity that are commonly accepted by humans. If the objective is to develop genuinely creative AI systems, it is worthwhile to consider whether they should replicate human creativity or instead explore novel forms of creativity that are uniquely enabled by their capabilities. The essence of AI creativity may lie not in emulating human creativity, but in making use of the distinctive capabilities of AI to generate outputs that are beyond the capabilities of humans to conceive or execute. This could have the effect of expanding the very definition of creativity itself. Some studies propose a method akin to the Turing test to assess whether an AI can generate creative outputs. If an AI can generate outcomes that mirror the works of a group of human artists, it demonstrates a level of creativity comparable to that group [117]. However, this capacity to emulate the training data does not entirely align with the conceptualisation of novelty as it pertains to the definition of creativity.

Possible implementation of the creative parameter

This raises the unanswered question of whether artificial intelligence must emulate the creative processes observed in humans in order to be considered a genuinely creative entity. This section will discuss the potential for creativity to become a possible parameter of this generative model, even without answering the question.

Fine-tuning of input text: The evaluation of the creativity of AI-generated images should consider both the artistic merit of the image and the extent to which it reflects the original textual prompt. The quality and specificity of the prompt provided to these models has a considerable impact on the creativity of the resulting output. It is therefore essential that the evaluation process takes into account not only the extent to which the image matches and extends the initial prompt, but also the degree of artistic creativity evident in the image itself. An additional step that can facilitate this creative process is natural language understanding (NLU). The introduction of an intermediate stage, during which the initial prompt is analysed and potentially reworked in order to enhance its creative potential, allows for the final output to be guided towards greater creativity. The NLU-based process could serve to refine the initial input, thereby emphasising creative aspects or introducing new perspectives. These could then be interpreted and visualised by the text-to-image model.

Evaluation through user feedback: As implemented in the high-fidelity playground model, the evaluation of a particular model through user evaluation has the potential to alter the dataset and foster greater creativity over time. This solution offers the potential for a unified approach to normal and highly creative architecture, allowing users to select the level of adherence to the required text and the degree of abstraction and artistic representation they desire. However, this approach may require a longer time frame for deployment and evolution, particularly if the initial model is not adequately prepared to accommodate the desired representation of creativity.

External semantic memory: The incorporation of an additional model representing semantic memory into architectural frameworks or the temperature for transformers may prove an effective means of enhancing the creative capacity of the underlying models. It is essential to consider the simplicity of comprehension, the degree of originality, and the extent of exploration of the semantic graph, as the criteria for evaluating originality may vary, as evidenced by the temperature parameter for transformers.

Conclusion: While text-to-image AI models rely on mathematical calculations, they are capable of generating images that may deviate from the orig-

inal imagination of the commissioning author. The debate surrounding AI creativity is contingent upon the definition of creativity itself. If creativity is defined as the capacity to produce something novel and useful, then it can be argued that AI models can be considered creative, albeit within certain constraints. These models are capable of generating new outputs, but not entirely new concepts or expressions.

The process of text-to-image generation is not merely an algorithmic execution; rather, it resembles a collaborative endeavour that is analogous to commissioned artwork. Although machines perform mathematical calculations, similar to heuristic methods employed by humans, if heuristics are considered a hallmark of human creativity, then machines could be deemed creative as well. This is due to the fact that they utilise a variety of techniques in order to achieve the final result. Although machines are fundamentally calculators, it is an incomplete perspective to dismiss them as uncreative. They are designed to produce accurate visual representations of input text; however, they are also capable of generating images without explicit prompts, provided that they are activated by human users. This specification is of paramount importance. In this context, the images are devoid of any subjective element, which gives the perceptions of randomness. Moreover, the quality of the prompt has a considerable effect on the final result. However, the definition of what constitutes a ”high-quality” prompt is dependent on the specific model in question. Consequently, the expression of human creativity in this process does not occur in a traditional sense but rather through the selection of precise language for the model and an understanding of its training data and configuration parameters. These factors are crucial for generating high-fidelity images [118], making prompt engineering a learned skill that requires expertise in crafting effective prompts and modifiers [119]. Nevertheless, the existence of hybrid AI architectures underscores the machines' lack of inherent creativity. They require substantial initial input to produce outputs that might be considered creative. This underscores the collaborative nature of the creative process between humans and machines. The analysis of AI-generated images, therefore, should consider the creativity embedded in the initial prompt, highlighting the symbiotic relationship between human input and machine output. In conclusion, while human and machine creativity differ, there is a compelling case for recognizing a distinct category of ”Artificial Creativity” [120]. This concept acknowledges the originality and effectiveness of AI-generated outputs while distinguishing them from human creativity. As noted by Google researchers in the Parti paper: ”Like a paintbrush, these models are a kind of tool that on their own do not produce art—instead, people use these tools to develop concepts and push their creative vision forward.”

References

[1] J. A. Plucker, R. A. Beghetto, and G. T. Dow, “Why isn't creativity more important to educational psychologists? potentials, pitfalls, and future directions in creativity research,” *Educational Psychologist*, vol. 39, no. 2, pp. 83–96, 2004. DOI: 10.1207/s15326985ep3902_1.

[2] S. Harvey and J. Berry, “Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity,” *Academy of Management Review*, 2022. DOI: 10.5465/amr.2020.0110.

[3] F. Magni, J. Park, and M. M. Chao, “Humans as creativity gatekeepers: Are we biased against AI creativity?” *Journal of Business and Psychology*, 2023, ISSN: 1573-353X. DOI: 10.1007/s10869-023-09910-x. [Online]. Available: https://doi.org/10.1007/s10869-023-09910-x.

[4] J. Lloyd-Cox, A. Pickering, and J. Bhattacharya, “Evaluating creativity: How idea context and rater personality affect considerations of novelty and usefulness,” *Creativity Research Journal*, vol. 34, no. 4, pp. 373–390, 2022. DOI: 10.1080/10400419.2022.2125721.

[5] A. Elgammal and B. Saleh, “Quantifying creativity in art networks,” Jun. 2015.

[6] M. Rhodes, "An analysis of creativity," *The Phi Delta Kappan*, vol. 42, no. 7, pp. 305–310, 1961, ISSN: 00317217. [Online]. Available: http://www.jstor.org/stable/20342603 (visited on 04/07/2024).

[7] G. E. Corazza, S. Agnoli, and S. Mastria, "The dynamic creativity framework," *European Psychologist*, vol. 27, no. 3, pp. 191–206, 2022. DOI: 10.1027/1016-9040/a000473.

[8] A. W. Flaherty, "Frontotemporal and dopaminergic control of idea generation and creative drive," *The Journal of Comparative Neurology*, vol. 493, no. 1, pp. 147–153, 2005. DOI: 10.1002/cne.20768. [Online]. Available: https://doi.org/10.1002/cne.20768.

[9] A. Mazza, O. Dal Monte, S. Schintu, *et al.*, "Beyond alpha-band: The neural correlate of creative thinking," *Neuropsychologia*, vol. 179, p. 108 446, 2023, ISSN: 0028-3932. DOI: https://doi.org/10.1016/j.neuropsychologia.2022.108446. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0028393222003050.

[10] J. S. Katz, M. R. Forloines, L. R. Strassberg, and B. Bondy, "Observational drawing in the brain: A longitudinal exploratory fmri study," *Neuropsychologia*, vol. 160, p. 107 960, 2021, ISSN: 0028-3932. DOI: https://doi.org/10.1016/j.neuropsychologia.2021.107960. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002839322100213X.

[11] R. E. Jung, J. M. Segall, H. Jeremy Bockholt, *et al.*, "Neuroanatomy of creativity," *Human Brain Mapping*, vol. 31, no. 3, pp. 398–409, 2010. DOI: https://doi.org/10.1002/hbm.20874. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20874. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20874.

[12] A. Schlegel, P. Alexander, S. V. Fogelson, *et al.*, "The artist emerges: Visual art learning alters neural structure and function," *NeuroImage*, vol. 105, pp. 440–451, 2015, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2014.11.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811914009318.

[13] R. K. Sawyer and D. Henriksen, *Explaining Creativity: The Science of Human Innovation*. Oxford University Press, Dec. 2023, ISBN: 9780197747537. DOI: 10.1093/oso/9780197747537.001.0001. [Online]. Available: https://doi.org/10.1093/oso/9780197747537.001.0001.

[14] A. Dietrich, "Types of creativity," *Psychonomic Bulletin & Review*, vol. 26, no. 1, pp. 1–12, 2019. DOI: 10.3758/s13423-018-1517-7. [Online]. Available: https://doi.org/10.3758/s13423-018-1517-7.

[15] J. P. Guilford, *The Nature of Human Intelligence*. McGraw-Hill, 1967.

[16] G. Wallas, *The Art of Thought*. London: Jonathan Cape, 1926.

[17] D. J. Treffinger, S. G. Isaksen, and K. B. Stead-Dorval, *Creative Problem Solving: An Introduction*, 4th. Routledge, 2006. DOI: 10.4324/9781003419327. [Online]. Available: https://doi.org/10.4324/9781003419327.

[18] J. D. Bransford and B. S. Stein, *The Ideal Problem Solver* (Book Library 46). Centers for Teaching Excellence, 1993.

[19] R. J. Sternberg, "The nature of creativity," *Creativity Research Journal*, vol. 18, no. 1, pp. 87–98, 2006, Retraction published 2020, Creativity Research Journal, 32(2), 200. DOI: 10.1207/s15326934crj1801_10. [Online]. Available: https://doi.org/10.1207/s15326934crj1801_10.

[20] P. Burnard, A. Craft, and T. Cremin, "Documenting 'possibility thinking': A journey of collaborative enquiry," *International Journal of Early Years Education*, vol. 14, Jan. 2006.

[21] S. Mednick, "The associative basis of the creative process," *Psychological Review*, vol. 69, no. 3, pp. 220–232, 1962. DOI: 10.1037/h0048850. [Online]. Available: https://doi.org/10.1037/h0048850.

[22] S. Denervaud, A. P. Christensen, Y. N. Kenett, and R. E. Beaty, "Education shapes the structure of semantic memory and impacts creative thinking," *NPJ Science of Learning*, vol. 6, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245013152.

[23] K. A. Ericsson, "The influence of experience and deliberate practice on the development of superior expert performance," in *The Cambridge Handbook of Expertise and Expert Performance*, K. A. Ericsson, N. Charness, P. J. Feltovich, and R. R. Hoffman, Eds., Cambridge University Press, 2006, pp. 683–703. DOI: 10.1017/CBO9780511816796.038. [Online]. Available: https://doi.org/10.1017/CBO9780511816796.038.

[24] A. S. Souza and L. C. Leal Barbosa, "Should we turn off the music? music with lyrics interferes with cognitive tasks," *Journal of cognition*, vol. 6, no. 1, p. 24, 2023. DOI: 10.5334/joc.273.

[25] K. R. Beittel and R. C. Burkhart, "Strategies of spontaneous, divergent, and academic art students," 1963. [Online]. Available: https://api.semanticscholar.org/CorpusID:151536100.

[26] J. W. Getzels, "Creative thinking, problem-solving, and instruction," *Teachers College Record*, vol. 65, no. 9, pp. 240–267, 1964. DOI: 10.1177/016146816406500910.

[27] C. S. Blair and M. D. Mumford, "Errors in idea evaluation: Preference for the unoriginal?" *The Journal of Creative Behavior*, vol. 41, pp. 197–222, 2007. DOI: 10.1002/j.2162-6057.2007.tb01288.x. [Online]. Available: https://doi.org/10.1002/j.2162-6057.2007.tb01288.x.

[28] K. Duncker, "A qualitative (experimental and theoretical) study of productive thinking (solving of comprehensible problems)," *Pedagogical Seminary and Journal of Genetic Psychology*, vol. 33, no. 4, pp. 642–708, 1926, ISSN: 0885-6559. DOI: 10.1080/08856559.1926.10533052.

[29] C. Salvi, E. Bricolo, J. Kounios, E. Bowden, and M. Beeman, "Insight solutions are correct more often than analytic solutions," *Thinking & Reasoning*, vol. 22, no. 4, pp. 443–460, 2016. DOI: 10.1080/13546783.2016.1141798.

[30] M. Benedek, T. Schües, R. E. Beaty, *et al.*, "To create or to recall original ideas: Brain processes associated with the imagination of novel object uses," *Cortex*, vol. 99, pp. 93–102, 2018, ISSN: 0010-9452. DOI: https://doi.org/10.1016/j.cortex.2017.10.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010945217303726.

[31] K. P. Madore, D. R. Addis, and D. L. Schacter, "Creativity and memory: Effects of an episodic-specificity induction on divergent thinking," *Psychological Science*, vol. 26, no. 9, pp. 1461–1468, 2015. DOI: 10.1177/0956797615591863. [Online]. Available: https://doi.org/10.1177/0956797615591863.

[32] M. Benedek, J. Jurisch, K. Koschutnig, A. Fink, and R. E. Beaty, "Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes," *NeuroImage*, vol. 210, p. 116586, 2020, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2020.116586. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920300732.

[33] W. Wan and C. Y. Chiu, "Effects of novel conceptual combination on creativity," *The Journal of Creative Behavior*, vol. 36, Dec. 2002. DOI: 10.1002/j.2162-6057.2002.tb01066.x.

[34] X. Mao, O. Galil, Q. Parrish, and C. Sen, "Evidence of cognitive chunking in free-hand sketching during design ideation," *Design Studies*, vol. 67, pp. 1–26, 2020. DOI: 10.1016/j.destud.2019.11.009.

[35] E. S. Joyce Gubbels and L. Verhoeven, "Predicting the development of analytical and creative abilities in upper elemen-

tary grades," *Creativity Research Journal*, vol. 29, no. 4, pp. 433–441, 2017. DOI: 10.1080/10400419.2017.1376548.

[36] G. Goldschmidt, "Visual displays for design: Imagery, analogy and databases of visual images," in *Visual Databases in Architecture*, A. Koutamanis, H. Timmermans, and I. Vermeulen, Eds., Avebury: Sage, 1995, pp. 53–74. DOI: 10.1080/10400410903579494.

[37] M. Ovando-Tellez, M. Benedek, Y. N. Kenett, *et al.*, "An investigation of the cognitive and neural correlates of semantic memory search related to creative ability," *Commun Biol*, vol. 5, p. 604, 2022. DOI: 10.1038/s42003-022-03547-x.

[38] Y. N. Kenett, D. Anaki, and M. Faust, "Investigating the structure of semantic networks in low and high creative persons," *Frontiers in Human Neuroscience*, vol. 8, 2014, ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00407. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00407.

[39] R.-Y. Horng, C.-W. Wang, Y. Yen, C.-Y. Lu, and C.-T. Li, "A behavioural measure of imagination based on conceptual combination theory," *Creativity Research Journal*, vol. 33, pp. 376–387, Oct. 2021. DOI: 10.1080/10400419.2021.1943136.

[40] J. A. Olson, J. Nahas, D. Chmoulevitch, S. J. Cropper, and M. E. Webb, "Naming unrelated words predicts creativity," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, e2022340118, 2021. DOI: 10.1073/pnas.2022340118.

[41] R. E. Beaty, A. P. Christensen, M. Benedek, P. J. Silvia, and D. L. Schacter, "Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production," *Neuroimage*, vol. 148, pp. 189–196, 2017. DOI: 10.1016/j.neuroimage.2017.01.012.

[42] R. E. Beaty, P. J. Silvia, E. C. Nusbaum, E. Jauk, and M. Benedek, "The roles of associative and executive processes in creative cognition," *Memory Cogn.*, vol. 42, pp. 1186–1197, 2014. DOI: 10.3758/s13421-014-0428-8.

[43] A. Fink and M. Benedek, "Eeg alpha power and creative ideation," *Neurosci. Biobehav. Rev.*, vol. 44, pp. 111–123, 2014. DOI: 10.1016/j.neubiorev.2012.12.002.

[44] T. Palmer, "Human creativity and consciousness: Unintended consequences of the brain's extraordinary energy efficiency?" *Entropy*, vol. 22, p. 281, 2020. DOI: 10.3390/e22030281.

[45] M. Benedek and A. Fink, "Toward a neurocognitive framework of creative cognition: The role of memory, attention, and cognitive control," *Current Opinion in Behavioral Sciences*, vol. 27, pp. 116–122, 2019, Creativity, ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2018.11.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352154618301839.

[46] E. Blaser, Z. W. Pylyshyn, and A. O. Holcombe, "Tracking an object through feature space," *Nature*, vol. 408, pp. 196–199, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:4418346.

[47] J. Duncan, "Selective attention and the organization of visual information," *Journal of Experimental Psychology: General*, vol. 113, no. 4, pp. 501–517, 1984. DOI: 10.1037/0096-3445.113.4.501. [Online]. Available: https://doi.org/10.1037/0096-3445.113.4.501.

[48] S. J. Luck, "Visual short-term memory," in *Visual Memory*, ser. Oxford Series in Visual Cognition, S. J. Luck and A. Hollingworth, Eds., Online edition, accessed 8 Apr. 2024, New York: Oxford University Press, 2008. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780195305487.003.0003.

[49] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature*, vol. 390, no. 6657, pp. 279–281, 1997. DOI: 10.1038/36846.

[Online]. Available: https://doi.org/10.1038/36846.

[50] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 8, pp. 2554–2558, 1982. DOI: 10.1073/pnas.79.8.2554. [Online]. Available: https://doi.org/10.1073/pnas.79.8.2554.

[51] D. Checiu, M. Bode, and R. Khalil, "Reconstructing creative thoughts: Hopfield neural networks," *Neurocomputing*, vol. 575, p. 127 324, 2024, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2024.127324. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122400095X.

[52] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, and S. Hochreiter, "Txt2img-mhn: Remote sensing image generation from text using modern hopfield networks," *IEEE Transactions on Image Processing*, vol. 32, pp. 5737–5750, 2023, ISSN: 1941-0042. DOI: 10.1109/tip.2023.3323799. [Online]. Available: http://dx.doi.org/10.1109/TIP.2023.3323799.

[53] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Addison-Wesley/Addison Wesley Longman, 1991.

[54] A. Dietrich and R. Kanso, "A review of EEG, ERP, and neuroimaging studies of creativity and insight," *Psychological Bulletin*, vol. 136, no. 5, pp. 822–848, 2010. DOI: 10.1037/a0019749. [Online]. Available: https://doi.org/10.1037/a0019749.

[55] M. Dinar, J. J. Shah, J. Cagan, L. Leifer, J. Linsey, S. R. Smith, *et al.*, "Empirical studies of designer thinking: Past, present, and future," *J. Mech. Des.*, vol. 137, p. 021 101, 2015. DOI: 10.1115/1.4029025.

[56] V. Goel and O. Vartanian, "Dissociating the roles of right ventral lateral and dorsal lateral prefrontal cortex in generation and maintenance of hypotheses in set-shift problems," *Cerebral Cortex*, vol. 15, no. 8, pp. 1170–1177, 2005. DOI: 10.1093/cercor/bhh217. [Online]. Available: https://doi.org/10.1093/cercor/bhh217.

[57] P. Howard-Jones, S. Blakemore, E. Samuel, I. Rummers, and G. Claxton, "Semantic divergence and creative story generation: An fMRI investigation," *Cognitive Brain Research*, vol. 25, pp. 240–250, 2005. DOI: 10.1016/j.cogbrainres.2005.05.013.

[58] F. Sieborger, E. Ferstl, and D. Y. von Cramon, "Making sense of nonsense: An fMRI study of task induced inference processes during discourse comprehension," *Brain Research*, vol. 1166, pp. 77–91, 2007. DOI: 10.1016/j.brainres.2007.05.079.

[59] P. Hansen, P. Azzopardi, P. Matthews, and J. Geake, *Neural correlates of "creative intelligence" an fMRI study of fluid analogies*, Poster session presented at the annual conference of the Society for Neuroscience, New Orleans, LA, Nov. 2008. [Online]. Available: http://www.brookes.ac.uk/.

[60] A. Fink, R. Grabner, M. Benedek, *et al.*, "The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI," *Human Brain Mapping*, vol. 30, pp. 734–748, 2009. DOI: 10.1002/hbm.20538.

[61] L. Aziz-Zadeh, S.-L. Liew, and F. Dandekar, "Exploring the neural correlates of visual creativity," *Social Cognitive and Affective Neuroscience*, vol. 8, no. 4, pp. 475–480, Feb. 2012, ISSN: 1749-5016. DOI: 10.1093/scan/nss021. eprint: https://academic.oup.com/scan/article-pdf/8/4/475/27107965/nss021.pdf. [Online]. Available: https://doi.org/10.1093/scan/nss021.

[62] Y. Yin, P. Wang, and P. R. N. Childs, "Understanding creativity process through electroencephalography measurement on creativity-related cognitive factors," *Frontiers in Neuroscience*, vol. 16, 2022, ISSN: 1662-453X. DOI: 10.3389/fnins.2022.951272. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.951272.

[63] R. E. Beaty, Q. Chen, A. P. Christensen, Y. N. Kenett, P. J. Silvia, M. Benedek, *et al.*, "Default network contributions to episodic and semantic processing during divergent creative thinking: A representational similarity analysis," *Neuroimage*, vol. 209, p. 116 499, 2020. DOI: 10.1016/j.neuroimage.2019.116499.

[64] A. Ali, R. Afridi, T. A. Soomro, S. A. Khan, M. Y. A. Khan, and B. S. Chowdhry, "A single-channel wireless eeg headset enabled neural activities analysis for mental healthcare applications," *Wireless Pers. Commun.*, vol. 125, pp. 3699–3713, 2022. DOI: 10.1007/s11277-022-09731-w.

[65] Y.-Y. Wang, T.-H. Weng, I.-F. Tsai, J.-Y. Kao, and Y.-S. Chang, "Effects of virtual reality on creativity performance and perceived immersion: A study of brain waves," *Br. J. Educ. Technol.*, pp. 1–22, 2022. DOI: 10.1111/bjet.13264.

[66] R. Sharpe and M. Mahmud, "Effect of the gamma entrainment frequency in pertinence to mood, memory and cognition," in *International Conference on Brain Informatics*, M. Mahmud, S. Vassanelli, M. S. Kaiser, and N. Zhong, Eds., Cham: Springer, 2020, pp. 50–61. DOI: 10.1007/978-3-030-59277-6_5.

[67] J. Bhattacharya and H. Petsche, "Shadows of artistry: Cortical synchrony during perception and imagery of visual art," *Cognitive Brain Research*, vol. 13, pp. 179–186, 2002. DOI: 10.1016/S0926-6410(01)00110-0.

[68] J. Bhattacharya and H. Petsche, "Drawing on mind's canvas: Differences in cortical integration patterns between artists and non-artists," *Human Brain Mapping*, vol. 26, pp. 1–14, 2005. DOI: 10.1002/hbm.20104.

[69] R. Solso, "Brain activities in a skilled versus a novice artist: An fMRI study," *Leonardo*, vol. 34, pp. 31–34, 2001. DOI: 10.1162/002409401300052479.

[70] J. Pearson, "The human imagination: The cognitive neuroscience of visual mental imagery," *Nature Reviews Neuroscience*, vol. 20, no. 10, pp. 624–634, 2019. DOI: 10.1038/s41583-019-0202-9. [Online]. Available: https://doi.org/10.1038/s41583-019-0202-9.

[71] F. J. Griffith and V. P. Bingman, "Drawing on the brain: An ale meta-analysis of functional brain activation during drawing," *The Arts in Psychotherapy*, vol. 71, p. 101 690, 2020, ISSN: 0197-4556. DOI: https://doi.org/10.1016/j.aip.2020.101690. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0197455620300630.

[72] M. Baas, C. K. De Dreu, and B. A. Nijstad, "A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus?" *Psychological bulletin*, vol. 134, no. 6, pp. 779–806, 2008. DOI: 10.1037/a0012815.

[73] M. Baas, B. A. Nijstad, and C. K. W. De Dreu, "Editorial: "the cognitive, emotional and neural correlates of creativity"," *Frontiers in Human Neuroscience*, vol. 9, 2015, ISSN: 1662-5161. DOI: 10.3389/fnhum.2015.00275. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00275.

[74] M. Roskes, C. De Dreu, and B. Nijstad, "Necessity is the mother of invention: Avoidance motivation stimulates creativity through cognitive effort," *Journal of personality and social psychology*, vol. 103, pp. 242–56, May 2012. DOI: 10.1037/a0028442.

[75] R. P. Heitz and J. D. Schall, "Neural mechanisms of speed-accuracy tradeoff," *Neuron*, vol. 76, no. 3, pp. 616–628, 2012. DOI: 10.1016/j.neuron.2012.08.030.

[76] A. Gidon, T. A. Zolnik, P. Fidzinski, *et al.*, "Dendritic action potentials and computation in human layer 2/3 cortical neurons," *Science*, vol. 367, no. 6473, pp. 83–87, 2020. DOI: 10.1126/science.aax6239. eprint: https://www.science.org/doi/pdf/10.1126/science.aax6239. [Online]. Available:

https://www.science.org/doi/abs/10.1126/science.aax6239.

[77] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton* (460-461). Cornell Aeronautical Laboratory, 1957, vol. 85.

[78] D. Beniaguev, I. Segev, and M. London, "Single cortical neurons as deep artificial neural networks," *Neuron*, vol. 109, no. 17, 2727–2739.e3, 2021, ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2021.07.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627321005018.

[79] L. Pessoa, "Understanding brain networks and brain organization," *Physics of Life Reviews*, vol. 11, no. 3, pp. 400–435, 2014. DOI: 10.1016/j.plrev.2014.03.005. [Online]. Available: https://doi.org/10.1016/j.plrev.2014.03.005.

[80] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 2014.

[81] M. Kang, J.-Y. Zhu, R. Zhang, *et al.*, *Scaling up GANs for text-to-image synthesis*, arXiv e-print, 2023. arXiv: 2303.05511 [cs.CV].

[82] P. Dhariwal and A. Nichol, *Diffusion models beat gans on image synthesis*, 2021. arXiv: 2105.05233 [cs.LG].

[83] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL].

[84] J. Yu, Y. Xu, J. Y. Koh, *et al.*, *Scaling autoregressive models for content-rich text-to-image generation*, 2022. arXiv: 2206.10789 [cs.CV].

[85] N. Geschwind, "The organization of language and the brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior.," *Science*, vol. 170, no. 3961, pp. 940–944, 1970.

[86] K. J. Holyoak, "Analogy and relational reasoning," in *The Oxford Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrison, Eds., Oxford University Press, 2012, pp. 234–259. DOI: 10.1093/oxfordhb/9780199734689.013.0013.

[87] L. Yu, B. Shi, R. Pasunuru, *et al.*, *Scaling autoregressive multi-modal models: Pretraining and instruction tuning*, 2023. arXiv: 2309.02591 [cs.LG].

[88] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. arXiv: 1503.03585 [cs.LG].

[89] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG].

[90] C. Saharia, W. Chan, S. Saxena, *et al.*, *Photorealistic text-to-image diffusion models with deep language understanding*, 2022. arXiv: 2205.11487 [cs.CV].

[91] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[92] T. Chen, *On the importance of noise scheduling for diffusion models*, 2023. arXiv: 2301.10972 [cs.CV].

[93] A. Nichol and P. Dhariwal, *Improved denoising diffusion probabilistic models*, 2021. arXiv: 2102.09672 [cs.LG].

[94] T. Karras, M. Aittala, T. Aila, and S. Laine, *Elucidating the design space of diffusion-based generative models*, 2022. arXiv: 2206.00364 [cs.CV].

[95] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV].

[96] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical text-conditional image generation with clip latents*, 2022. arXiv: 2204.06125 [cs.CV].

[97] J. Betker, G. Goh, L. Jing, *et al.*, "Improving image generation with better captions." [Online]. Available: https://api.semanticscholar.org/CorpusID:264403242.

[98] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, *Glide: Towards photorealistic image generation and editing with text-guided diffusion models*, 2022. arXiv: 2112.10741 [cs.CV].

[99] C. Mihaly, "Artistic problems and their solutions: An exploration of creativity in the arts," Ph.D. dissertation, University of Chicago, 1965.

[100] P. Esser, S. Kulal, A. Blattmann, *et al.*, *Scaling rectified flow transformers for high-resolution image synthesis*, 2024. arXiv: 2403.03206 [cs.CV].

[101] D. Podell, Z. English, K. Lacey, *et al.*, *Sdxl: Improving latent diffusion models for high-resolution image synthesis*, 2023. arXiv: 2307.01952 [cs.CV].

[102] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, Mar. 2021, ISSN: 1573-1405. DOI: 10.1007/s11263-021-01453-z. [Online]. Available: http://dx.doi.org/10.1007/s11263-021-01453-z.

[103] C. Meng, R. Rombach, R. Gao, *et al.*, *On distillation of guided diffusion models*, 2023. arXiv: 2210.03142 [cs.CV].

[104] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi, *Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation*, 2024. arXiv: 2402.17245 [cs.CV].

[105] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, *Make-a-scene: Scene-based text-to-image generation with human priors*, 2022. arXiv: 2203.13131 [cs.CV].

[106] K. Sohn, N. Ruiz, K. Lee, *et al.*, *Styledrop: Text-to-image generation in any style*, 2023. arXiv: 2306.00983 [cs.CV].

[107] H. Chang, H. Zhang, J. Barber, *et al.*, *Muse: Text-to-image generation via masked generative transformers*, 2023. arXiv: 2301.00704 [cs.CV].

[108] S. Barratt and R. Sharma, *A note on the inception score*, 2018. arXiv: 1801.01973 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1801.01973.

[109] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982, ISSN: 0047-259X. DOI: https://doi.org/10.1016/0047-259X(82)90077-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0047259X8290077X.

[110] M. J. Chong and D. Forsyth, *Effectively unbiased fid and inception score and where to find them*, 2020. arXiv: 1911.07023 [cs.CV].

[111] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, *Rethinking fid: Towards a better evaluation metric for image generation*, 2024. arXiv: 2401.09603 [cs.CV].

[112] J. Li, D. Li, C. Xiong, and S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022. arXiv: 2201.12086 [cs.CV].

[113] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, *The unreasonable effectiveness of deep features as a perceptual metric*, 2018. arXiv: 1801.03924 [cs.CV].

[114] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, *Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size*, 2016. arXiv: 1602.07360 [cs.CV].

[115] J. Xu, X. Liu, Y. Wu, *et al.*, *Imagereward: Learning and evaluating human preferences for text-to-image generation*, 2023. arXiv: 2304.05977 [cs.CV].

[116] D. Foster, *Generative Deep Learning*. O'Reilly Media, Inc., 2022.

[117] H. Wang, J. Zou, M. Mozer, *et al.*, *Can ai be as creative as humans?* 2024. arXiv: 2401.01623 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2401.01623.

[118] J. Oppenlaender, "The creativity of text-to-image generation," in *Proceedings of the 25th International Academic Mindtrek*

Conference, ser. Academic Mindtrek 2022, ACM, Nov. 2022. DOI: 10.1145/3569219.3569352. [Online]. Available: http://dx.doi.org/10.1145/3569219.3569352.

[119] J. Oppenlaender, "A taxonomy of prompt modifiers for text-to-image generation," *Behaviour & Information Technology*, Nov. 2023, ISSN: 1362-3001. DOI: 10.1080/0144929x.2023.2286532. [Online]. Available: http://dx.doi.org/10.1080/0144929X.2023.2286532.

[120] M. A. Runco, "Ai can only produce artificial creativity," *Journal of Creativity*, vol. 33, no. 3, p. 100 063, 2023, ISSN: 2713-3745. DOI: https://doi.org/10.1016/j.yjoc.2023.100063. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2713374523000225.