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

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

June 2, 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. 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 aim is to identify similarities and differences between the two systems. 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 various domains, such as image generation and natural language processing, replicating human creativity remains a challenge. This may be attributed to the elusive nature of creativity, which defies straightforward definition and quantification. Alternatively, when evaluating creativity, it is crucial to avoid biases that may be induced by the characteristics of the creator. It is often assumed that generative AI produces less effort than humans in creating a given artefact, which results 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, specifically when AI generates visual content based on human textual input. The objective of this paper is to clarify our comprehension of human cognition and AI by examining the similarities and distinctions between human and AI-driven creativity. This paper will examine the neural architecture underlying human visual creativity and its AI counterparts.

Defining Creativity: Assessing creativity remains a contentious subject in the creativity research. Numerous frameworks and assessment tools have been devised over time to establish standardised definitions and metrics for creativity. 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, there has been a growing consensus around a definition that appears to link two main concepts. Creativity is defined as the production of something both novel and useful within a given social context [8]. However, it remains difficult to execute scientific tests in a rigorous way. Indeed, most of the tests analysed in this review were judged by human beings, independent judges with objectivity but each with a different perception of significance [9]–[12]. Psychometricians emphasise the importance of validity and reliability in the evaluation of creative abilities [13]. Validity and reliability serve as principles in the researches. A test is considered valid if it effectively captures the intended aspect of creativity. Conversely, reliability refers to the consistency of results obtained when the same test is administered to the same individual on multiple occasions.

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 each 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]. This process has been the subject of multiple studies, which have demonstrated that creativity tends to occur in a sequence of more stages [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 related 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. As you notice, the conscious mind plays an important role in the creative process. 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. 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 [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. The process, which is established within a committee, is agreed with the contractor explicitly or implicitly through feedback on the final work.

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. Studies examine semantic memory search processes that support higher creative thinking, identifying clustering as correlated with divergent thinking and switching with the ability to combine distant associates. The results suggest that clustering is more associated with divergent thinking, while switching is associated with convergent thinking and a more efficient 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].

Brain flow and structure: Creativity is not specifically associated with any single area of the brain [50]. 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 fMRI [51]. Results from numerous fMRI studies consistently highlight the central role of the prefrontal cortex in both hemispheres [52]–[56], presumably due to its involvement in working memory and executive attention processes. This observations aligns with the framework of associative 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, functional magnetic resonance imaging (fMRI) was used by researchers to evaluate neural activity in individuals during a visuospatial creativity task [57]. 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 (EEG) data with existing research on brain activation during creative cognitive tasks. It was found that remote association is associated with frontal lobe activity [58]. This contrasts with some previous studies that associate it with temporal lobe activity [59]. 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 [60], theta [61], and gamma waves [62]. 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 [63]–[65]. Divergent thinking is commonly measured using the Alternative Uses Test (AUT) [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 [66]. 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]. This highlights the importance of sensory and motor experiences in promoting creativity. 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 [67].

III. Artificial Neural networks

As explained in the Human Architecture section, the creative process relies heavily on semantic memory. However, most current architectures do not incorporate explicitly a network that emulates the human memorisation process, despite its importance. The Hopfield Neural network (HNN) [68] is know to his capacity to emulate human memory assotiation process. Recent study introduces the use of them as a tool to emulate the creative process through concept association [69]. 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 [70]. 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 [71]. Therefore, in the field of text-to-image synthesis, the three main methodologies - Generative Adversarial Networks (GANs), autoregressive methods, and diffusion models - do not explicitly use the HNN architecture.

Generative Adversarial Networks

Generative adversarial networks (GANs) [72] 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 [73], diffusion models have shown significant results that surpass the quality of GAN models [74].

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 [75]. This sequential generation differs markedly from the parallel approach employed by Generative Adversarial Networks (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.

Parti: Parti [76] 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. Increased prompt complexity may lead to errors such as color bleeding, omission, hallucination, duplication of details, displaced positioning or interactions. To prevent these errors, it is important to increase proportionally the number of tokens to generate a well-structured output.

CM3Leon: Meta has developed CM3Leon [77], a transformer-based autoregressive model. The aim of this approach is to balance inference time with global image coherence. This is accomplished through retrieval augmented pretraining on a large, diverse 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 uses a decoder-only transformer architecture without an encoder. To merge the two token vocabularies in the decoder, a break token is used to indicate when the text tokens stop and the image tokens start. The model is trained in two stages: pretraining and supervised fine-tuning. Contrastive decoding is used to improve sample quality, and the text is replaced with the mask token from the CM3 objective to enable unconditional sampling. The CM3 objective accepts multi-modal inputs, making it a versatile model capable of infilling and autoregressive generation tasks for both images and text. This allows for guidance without the need for finetuning. CM3Leon is fine-tuned on a wide array of mixed image and text tasks, which are organized as a series of interleaved text and image examples. Notably, the model has been trained exclusively on licensed images. It is important to note that other models may not have the same policy on training.

Diffusion Models architectures

Diffusion models [78], [79] have become popular in image generation due to their strong performance and relatively low computational cost [80]. 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 [81]. Although initially developed for the purpose of brain segmentation in the medical field, this convolutional architecture has since been demonstrated to be highly effective in a range of computer vision tasks. At the time of inference, by reversing the forward paths of data towards noise, it becomes feasible to generate data from noise. The noise predictor estimates the noise of the image, which is then subtracted from the image. This process is not conducted in a single instance, as it 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 employed is known as the sampler. The noise schedule regulates the level of noise at each sampling step, with the highest noise occurring at the first step and gradually decreasing to zero at the final step. The sampler's objective at each step is to generate an image with a noise level that corresponds to the noise schedule. Multiple noise samplers exist, and the optimal one depends on the task [82]. For istance, according to some study, a linear schedule is not well-suited for low-resolution images [83]. The samplers labelled 'Karras' use the noise schedule recommended in the Karras article[84]. 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. 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 [85], and it offers a significant improvement over the basic diffusion model. One of its key advantages is speed, as it is many times faster than denoising the raw, uncompressed data directly. The current dominant paradigms, diffusion models, and autoregressive models, both rely on iterative inference. While iterative methods facilitate stable training with straightforward objectives, they come with a significant computational overhead during inference.

CLIP: Before discussing OpenAI's models, it would be beneficial to first examine the CLIP technology. This technology enabled the company to make improvements to the text to image field. Although CLIP is not implemented in DALL-E 2 [86] and DALL-E 3 [87], it is present in other diffusion models. However, it should be noted that the two OpenAI models are based on it. CLIP, or Contrastive Language-Image Pre-training, is distinguished by its ability to comprehend the intricate relationship between images and text. This model comprises both an image encoder and a text encoder, which were trained on a dataset of 400 million images. The objective is to generate similar embeddings for images and captions that match, while producing distinctly different embeddings for those that do not match. This enables the model to effectively understand and relate images to their corresponding captions. CLIP employs a convolutional neural network (CNN) as its vision encoder for image processing. The CNN extracts high-level features, which are then refined through subsequent layers. On the textual side, CLIP employs a transformer-based text encoder that captures and encodes semantic information in a format compatible with image representations. In CLIP, both images and text are processed concurrently, allowing the model to learn to establish connections between image representations and their corresponding textual descriptions. To guarantee comparability across modalities, CLIP includes normalisation and projection layers. It employs a contrastive learning objective, which encourages connections between similar image-text pairs while distinguishing dissimilar ones. This promotes meaningful associations between images and text.

DALL-E 2: OpenAI employed this 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 [88] architecture. 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. The autoregressive approach employs principal components analysis (PCA) to reduce the image embeddings to a lower dimension, 319 dimensions. This not only optimises the mathematical process but also improves training stability. This reduced embedding pass is incorporated into a classical autoencoder of transformer, which outputs the new image embedding. In contrast, the diffusion process provides the transformer with the encoded text, the CLIP embedding, the noisy embedding with the relative timestamp, and returns the new image embedding as the output. In order to achieve this, the researchers employed the cosine noise schedule [83]. 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 [88], 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

DALL-E 3: DALL-E 3 [87] 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.

Imagen: Imagen is the diffusion model developed by Google [80]. For further details, please refer to the cited source. Architectural is similar to the latent diffusion model of stable. Its principal innovation is that it employs a substantial pretrained NLP model called T5-XXL, which is taken 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 encounters difficulties with regard to 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. The text-to-image diffusion model is based on the U-Net architecture and is adapted for 64×64 resolution images. This model is conditioned on text embeddings using a combined pooled embedding vector and diffusion timestep embedding. Furthermore, cross-attention is employed over text embeddings at different resolutions, and layer normalisation is incorporated to increase the performance of text embeddings.

Stable Diffusion: Stable Diffusion [85] is a model introduced by StabilityAI in 2022. It was the first model to have the ability to generate images based on specific text prompts or other conditioning inputs. This is achieved by introducing conditioning mechanisms into the inner diffusion model, which has also been referred to in the literature as Classifier-Free Guidance (CFG) [89]. The model takes an image and sends it to a variational autoencoder, where it is processed with diffusion in the latent space. In fact, the model was designated the Latent Diffusion Model (LDM) and was the inaugural system to demonstrate the potential for computational savings in the latent space. To facilitate conditioning, the denoising U-Net of the inner diffusion model employs a cross-attention mechanism, which enables the model to process inputs in a more efficient manner. The model has been improved, resulting in the third version, which incorporates the Rectified flow [90]. This is a generative model formulation that establishes a direct connection between data and noise. It employs a stochastic differential equation (SDE) as a sampling method. The rationale behind this approach is that SDE facilitates a transition from data distribution to noise distribution. The architectural framework comprises the following steps: the caption is taken, multiple encoders are employed to encode the information, and the resulting outputs are combined to construct an intermediate representation, which is then concatenated with the three outputs. These three encoders are two types of CLIP model and T5. Removing T5 has no effect on aesthetic quality ratings, and only a small impact on prompt adherence. However, its contribution to the capabilities of generating written text is more significant. The model's robustness is increased, and the issues associated with CLIP are mitigated. Furthermore, a pooled vector is employed to store the same information in a more lightweight manner. This representation, in conjunction with the sinusoidal encoding of time, represents high-level text information in time. The text representations and the latent representations are then processed with transformers, undergoing multiple cross-attention operations. The model was trained on an open-source dataset comprising unlabeled images, which were recaptioned in a manner similar to that employed by DALL-E 3 [87].

SDXL: This is another StabilityAI models architecture [91] that tries to combine the sample quality of diffusion models with the inherent speed of GANs. The development of this architecture commenced with a variational autoencoder, where it was processed with diffusion in the latent space. In the latent space, the image underwent two UNets, designated as base and refined, and was then decoded by the variational decoder. The base UNet executed the standard procedure, whereas the refined UNets introduced noise into the output of the first UNet, with fewer steps. This allows the model to raise the quality of the initial image produced, as the features are not altered to the same extent. Furthermore, they have resolved a significant shortcoming of their LDM paradigm, namely that training a model necessitates a minimum image size due to its two-step architecture. Rather than discarding the images and losing information, or upsampling the low-resolution images and introducing artefacts, they have devised a solution whereby the UNets can accept the image's dimensionality as input. Similarly, the image resolution of the final result can be augmented by incorporating height and weight. Furthermore, in comparison with the initial versions of Stable Diffusion, two encoders have been added and concatenated. As improvement, they proposed a novel technique called Adversarial Diffusion Distillation (ADD) [92], capable of reduce the overhead of inference in pre-trained diffusion models. ADD is a method of distilling pre-trained diffusion models into high-fidelity. The method achieves this by condensing the typically multi-step sampling process of diffusion models into just 1-4 steps while ensuring high sampling fidelity and potentially raising overall model performance. This technique leverages score distillation [93] to use large-scale off-the-shelf image diffusion models as a teacher signal combined with an adversarial loss to ensure high image fidelity even in scenarios with one or two sampling steps. The ADD student is trained as a denoiser that receives diffused input images and outputs samples. It is trained to improve two objectives: an adversarial loss and a distillation loss. The adversarial loss is designed to fool a discriminator and the distillation loss is designed to match the denoised targets of a frozen diffusion model teacher. Firstly, the adversarial loss instructs the model to generate samples that are directly aligned with the manifold of real images during each forward pass. This directive minimises the occurrence of undesirable artefacts such as blurriness, which is commonly encountered in other distillation methods [94]. Secondly, the distillation loss utilises the knowledge encoded in a separate, pretrained diffusion model, which acts as a fixed teacher. By distilling this knowledge into the student model, ADD capitalises on the extensive learned representations of the diffusion model, thereby preserving the robust compositionality characteristic of large-scale diffusion models. This streamlined process enables the efficient generation of high-quality samples with reduced computational overhead. The model's limitations include its inability to generate a pure black image or a pure white image, as well as its inability to place subjects onto solid backgrounds [82].

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 is a Latent Diffusion Model that uses two fixed, pre-trained text encoders (OpenCLIP-ViT/G and CLIP-ViT/L) and follows the same architecture as SDXL. The latest version of the model, Playground v2.5 [95], offers three insights that aim to improve the aesthetic quality of text-to-image generative models. The first insight concerns the significance of the noise schedule in training a diffusion model, which has a profound impact on realism and visual fidelity. The new model, in fact, changed the noise schedule using the Ediffusion model noise schedule [84]. The second insight addresses the challenge of accommodating various aspect ratios in image generation. It emphasises the importance of preparing a balanced bucketed dataset. The third insight investigates the crucial role of aligning model outputs with human preferences. This ensures that generated images resonate with human perceptual expectations. To archive this, they developed a system that enables them to automatically curate a high-quality dataset from multiple sources via their platform users' ratings. 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.

Hybrid architectures

Meta make a scene method: Meta's approach to image generation is designed to increase accuracy and relevance through the use of meta-feedback. This approach enables users to sketch their desired outputs and refine generated images. It was developed in response to the recognition that, despite recent advances, challenges persist in the quality and applicability of existing text-to-image models. Meta's method integrates various features into the classical autoregressive paradigm in order to address these challenges effectively. Firstly, Meta introduces a control mechanism, scene layout, which complements textual input, improving structural consistency and quality while enabling scene editing. The utilisation 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 towards generating images that align better with human preferences, particularly focusing on aspects such as faces and salient objects. text editing with anchor scenes, and overcoming out-of-distribution text prompts. Secondly, the tokenisation process is refined by incorporating domain-specific knowledge over crucial image regions, such as faces and salient objects. This enhances the overall representation of the token space and improves generation quality and alignment with textual input. A CFG mechanism is introduced to raise generation quality without relying on post-generation filtering. This allows the model to leverage implicit feedback during training, resulting in improved image fidelity and text alignment.

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 [96] uses a transformer-based model for text-to-image generation, specifically leveraging Muse, [97] 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. Adapter tuning is the technique utilized in Styledrop to style-tune the text-to-image transformer. The process entails creating a text input based on a style reference image. This input comprises content and style text descriptors, which have been designed to promote content-style disentanglement, a crucial factor in achieving compositional image synthesis. Styledrop employs an iterative training framework to rais model performance over successive iterations. This framework involves training a new adapter on images sampled from a previously trained adapter. This iterative approach enables the model to gradually enhance its comprehension of the relationships between style and content, resulting in improved synthesis outcomes. A CFG mechanism is employed to refine the generation process, thereby obviating the necessity for external classifiers or post-processing techniques. This further improves image quality and coherence.

IV. Discussion

Differences and analogies

Structural differences: AI 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 process as we have already discussed. 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 creative process. Conversely, relaxation, anger, and anxiety have a negative influence [98]–[100]. 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 [101]. It is impossible that any biological brain could implement the numerical precision required by ML in a sufficiently rapid manner to be useful. 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 [102], which perceptrons [103] 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 [104]. This implies that a single neuron in our body can be compared to an entire network. Interestingly, when the upper bounds of dendrites is not thresholded, the number of layers decreases to one.

Training process: 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 [105], which requires some modification to the basic perceptron algorithm. This may prevent backpropagation from working at all. In particular, the process of an artist learning by creating an image and then receiving feedback is more similar to Gans' process than diffusion models. The diffusion model process of reversing from an initial paint to a white canvas, which is the most analogous to the initial noise, is never undertaken. Moreover, as has been demonstrated, the quality of the training data affects the quality of the output. 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.

Memory: In human cognitive frameworks, knowledge representation is typically manifested by conceptualising entities as objects with attributes, encapsulating a hierarchical structure that reflects real-world semantics. In contrast, in artificial neural networks, particularly in our context of text-to-image processing, the intrinsic meaning associated with a word or concept is replaced by its correlation with visual representations. CLIP functionality appears to mirror the human ability to associate words with their mental representations, with the technique being employed in the majority of state-of-the-art diffusion model architectures. Although knowledge is organised in a structured manner, with concepts being related to one another by proximity and therefore aligned with associative theory, the CLIP embedding process itself fails to explicitly bind attributes to objects [86]. This produces limitations in the replication of attributes pertaining to subjects in the output and highlights the distinction between the our semantic memory and the CLIP counterpart. Moreover, 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. In contrast, CM3Leon employs a decoder transformer to comprehend the meaning of both modalities. This approach enables not only text-to-image conversion but also image-to-text conversion. This method bears resemblance to the human capacity to convert one modality to another. In fact, the transformer decoder accepts either text or image as input, with the vocabulary being merged. The mask system enables the user to select which modality is input and which is output. Although the CM3Leon architecture is similar, it does not represent a perfect reverse process of the human brain's flow.

Drawing process: In general, diffusion models are said to emulate the principle of the human process of sketching and refining the result. 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

[106]. Therefore, diffusion models exhibit similarities with the human brain, yet also exhibit critical differences in terms of the dimensions sampled. While the UNet pooling technique can be useful for understanding the meaning of an image, the latent space sampling approach is a computational-saving trick and is therefore not feasible for humans. Additionally, some models have a fixed width and height production of the image, which contrasts with the capacity of humans to adapt their representation according to their imagination.

Evaluation metrics

FID: In the term usufulness on the creativty definition we implicitly say that the image should be meaningful. Hallucinations or random noise have novelty but definitely not usufulness. FID is a metric that quantifies the degree of similarity between two datasets of images. It is employed in the assessment of the quality of generated images, evaluating their visual fidelity and diversity in comparison to the real images. The measure is calculated by computing the Fréchet distance [107] between two Gaussians fitted to feature representations of the Inception network. Despite its extensive utilisation within the industry, it has been demonstrated that FID may diverge from the assessments of human raters, particularly in significant cases [108]. It is evident that human raters represent the most valuable evaluators in this context. Consequently, it is inadvisable to rely solely on FID as a measurement.

LPIPS: Learned Perceptual Image Patch Similarity (LPIPS) [109] is designed to better capture human perception of image similarity compared to other metrics such as FID. The measure focuses on capturing high-level semantic similarity between images. It uses deep learning models such as SqueezeNet [110] to extract deep features from images, and then compares the distances between these deep features to measure perceptual similarity. Although LPIPS is trained on human judgments of perceptual similarity, it is vulnerable to adversarial attacks that can fool its neural network and thus its final judgement.

CMMD: Instead of estimating the Frechet distance, Google researchers propose using an alternative new metric called CMMD [111]. This metric is based on richer CLIP embeddings and the maximum mean discrepancy distance with the Gaussian RBF kernel. It is an unbiased estimator that does not make any assumptions on the probability distribution of the embeddings and is sample efficient. Image Reward simulate the commissioning of an image to an artist by integrating human preference feedback to enhance text-to-image models. After the generation phase, humans vote on four parameters: Overall satisfaction, adherence to the prompt, aesthetics, and meta-feedbacks.

Text-image scoring methods: Text-image scoring methods can assess whether a generated image matches a text prompt. Examples are CLIP itself or Bootstrapping Language-Image Pre-training (BLIP) [112]. BLIP uses a multimodal encoder-decoder architecture, with a unimodal encoder to reconcile visual and linguistic representations, and an image-based text encoder and decoder, which have a variety of attentional layers to meticulously assess compliance with the prompt. However, these methods do not always match human preferences and perceptions.

ImageReward: ImageReward [113] is a general-purpose text-to-image human preference reward model that encodes human preferences in the diffusion process. It outperforms existing scoring models and metrics in human evaluation, making it a promising automatic metric for evaluating text-to-image synthesis. ImageReward has been trained with 137,000 human scores of AI images and is expected to provide significantly better image syntheses then BLIP or CLIP. It introduces Reward Feedback Learning (ReFL) to tune diffusion models with respect to human preference scorers. This information about the quality in the final denoising steps allows direct feedback learning on diffusion models that do not provide a probability for their generations.

Conclusion: The need to increase the refinement of the input does not detract from the fact that for all intents and purposes, despite being a mathematical calculation, the AI is able to synthesize an image that differs from the commissioning author's imagination. This is for all intents and purposes assess-able as creativity. Nonetheless, hybrid architectures emphasise the importance of a holistic approach to evaluating the end result. As previously stated, the quality of the prompt exerts a significant influence on the final result. However, the definition of prompt quality is specific to the model in question. For this reason, human creativity is not expressed in the conventional manner; rather, it is manifested in the selection of an appropriate word selection for a given model and an understanding of the system's training data and configuration parameters. These factors are crucial for the generation of high-fidelity images [114]. The art of prompt engineering is a learned skill, as it is not immediately apparent how to write effective prompts and which keywords make good prompt modifiers [115]. The outcome of this debate depends on how we define and conceive creativity. If we adhere to the definition of novelty and usefulness, AI models can be considered creative. Nevertheless, the existence of these hybrid architectures highlights the lack of inherent creativity in machines, which require more initial information to produce something more creative. This text emphasises the collaborative nature of creative processes. The analysis of AI-generated images is evaluated in relation to the creativity of the initial prompt, thereby highlighting the symbiotic relationship between human input and machine output. The text-to-image process extends beyond algorithmic execution to embody a collaborative endeavour akin to commissioned artwork. The machine performs mathematical calculations, which is also done heuristically by humans. However, if we consider heuristics as the defining characteristic of human creativity, then machines can also be considered creative since they have different ways of arriving at the final output. It is important to acknowledge that machines are primarily calculators, but the hypothesis that they are not creative is incomplete. Despite the objective aim of text-to-image models to produce accurate visual representations of the input text, focusing on fidelity to the description rather than subjective interpretation, it is possible for these models to produce such images, without prompt, if activated by humans. In this context, we do not search for subjectiveness inside this picture because we are aware that it does not exist, and therefore it feels like mere randomness. In extent, we can argue that the two are different and then should exist the term Artificial creativity [116] that recognize the originality and effectiveness but also the differences with the human creativity. As the Google Researchers in the Parti paper posited: ”Like a paint brush, 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 metatheory 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. 116 586, 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 freehand 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 elementary 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] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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/.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] J. Bhattacharya and H. Petsche, "Shadows of artistry: Cortical synchrony during perception and imagery of visual art," *Cogni-*

tive Brain Research, vol. 13, pp. 179–186, 2002. DOI: 10.1016/S0926-6410(01)00110-0.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

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

[73] 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].

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

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

[76] 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].

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

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

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

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

[81] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical*

image segmentation, 2015. arXiv: 1505.04597 [cs.CV].

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

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

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

[85] 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].

[86] 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].

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

[88] 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].

[89] J. Ho and T. Salimans, *Classifier-free diffusion guidance*, 2022. arXiv: 2207.12598 [cs.LG].

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

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

[92] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, *Adversarial diffusion distillation*, 2023. arXiv: 2311.17042 [cs.CV].

[93] 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.

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

[95] 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].

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

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] A. Gidon, T. A. Zolnik, P. Fidzinski, *et al.*, "Dendritic action potentials and computa-

tion 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.

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

[104] 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.

[105] 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.

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

[107] 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.

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

[109] 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].

[110] 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].

[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] 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].

[114] 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.

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

[116] M. A. Runco, "Ai can only produce artificial creativity," *Journal of Creativity*, vol. 33, no. 3, p. 100063, 2023, ISSN: 2713-3745. DOI: https://doi.org/10.1016/j.yjoc.2023.100063. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2713374523000225.