Semi-Automated Evaluation of Visual Anagrams: Assessing Perception in Diffusion Models and Humans

Flavia-Diana Macovei

Department of Computer Science, University of Toronto

Abstract—Computer-generated optical illusions represent an intriguing area of research, reaching beyond straightforward image generation and accordingly providing a basis for the analysis of perception in generative models. This study aims to explore the abilities of diffusion models to produce artwork in the form of illusions and reveal correlations between machine-generated and human-made art. A semi-automated evaluation pipeline preprocesses illusions generated from a pre-trained diffusion model at a large scale, followed by a manual qualitative analysis. Additionally, the conceptual and computational limitations of the existing approach are explored through the implementation of additional transformation types and an ablation study of parallel denoising.

Index Terms—Computational Imaging, Diffusion Models, Perception, Computer-Generated Illusions

1 Introduction

Recent years have seen a considerable upturn in the use of text-to-image generation with popular models such as Imagen and DALL-E seeing widespread application [1–6]. Images synthesised from a natural language input can be utilised for data visualisation, image editing, and evidently for creative expression or "AI art". This growing demand for machine-based image generation raises questions regarding its similarity to human-made art as well as ethical implications of its application [4].

Diffusion models are a subtype of image generation models that produce an image by starting with a sample from a Gaussian noise distribution and iteratively denoising it to approximate an instance from the target distribution [7, 8]. The focus of this study is to examine the relationship between perception in diffusion models and human perception.

The task of clear-cut image generation has been subject to extensive research and noteworthy results have been accomplished in this field [9]. A more rigorous test of the capabilities of diffusion models lies in the generation of optical illusions, which engage with the viewer's perception and more closely resemble the inspired talent of an artist [10].

Geng et al. propose a methodology for creating optical illusions using a pre-trained diffusion model. Their approach involves estimating the noise corresponding to two or more target images under specified image transformations and subsequently averaging these estimates. Resulting images depict different targets depending on the transformation from which they are viewed [7]. These illusions present an opportunity for analysing the differences and similarities of perception in humans and artificial intelligence.

Besides evaluating the analogy of AI art and human art, this study aims to explore the boundaries of applicability and performance of the method proposed by Geng et al.

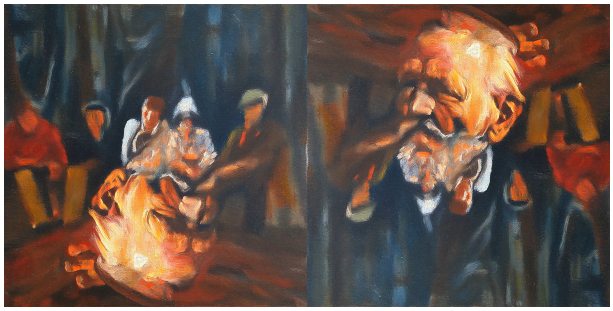


Fig. 1: Example of a visual anagram. Input configuration consists of targets *people at a campfire* and *an old man* with style *oil painting* under transformations *identity* and *flip*.

One such exploration is the addition of elaborate transformations. For this purpose, two new transformations, a colour channel permutation and a non-cardinal rotation, are implemented and probed.

The investigation of performance limits involves modifying the model architecture so that denoising is directed towards a single target at each step, with targets alternating, rather than averaging parallel noise estimates. This aspect is addressed in the original work, but the present study aims to further explore the effect of this modification on time complexity and quality of the resulting illusion.[1]

2 Related Work

Visual anagrams or multi-view optical illusions is the name that is given by Geng et al. to images which depict different targets depending on the transformation under which they are viewed [7] (see Figure 1 for an example using

1. The source code is publicly available at https://github.com/flaviamacovei/visual_anagrams

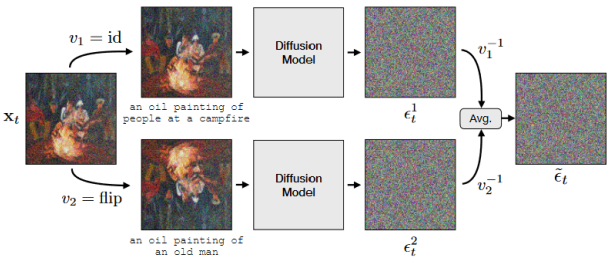


Fig. 2: Visual anagrams algorithm. All transformations are denoised simultaneously and their estimates averaged, resulting in a unified noise estimate at each inference step.

the *identity* and *flip* transformations). The authors present a model that generates these illusions by leveraging an adapted pre-trained diffusion model. At each inference step, the model evaluates the noise within the context of every specified transformation. These individual estimations are then inverted back to identity and combined through averaging, yielding a unified result for that step. Repetition of this parallel denoising method results in an image that resembles each specified target under the corresponding transformation.

Transformations provided by the authors include rotations and flips, skews, so-called poly-morphic jigsaws with more than one solution, and image negation. For a transformation to be deemed valid within this framework, it must satisfy the criteria of linearity and statistical consistency. That is to say, if a transformation $v(.)$ of an image x can be expressed as Ax for an orthogonal matrix A, the transformation is suitable.

The authors demonstrate that this model excels in producing high-quality illusions across a diverse range of transformations. One plausible cause is a fundamental similarity of perception in generative models and humans. It is stated that "generative models may process optical illusions in a way similar to humans", drawing on research performed on convolutional neural networks [10], generative classifiers [11], and large vision-language models [12].

From a quantitative standpoint, the success of an illusion is measured with the so-called *alignment* and *concealment* scores derived from the score matrix provided by CLIP [7]. CLIP or Contrastive Language-Image Pre-training is a method for image representation learning which at inference time can be used for predicting the probability of specified captions given an image [13, 14]. In the present context it is used to compute a score matrix $\mathrm{S} \in \mathbb{R}^{N \times N}$ where each entry represents the normalised probability of a particular caption (or target) belonging to a particular image (or transformation):

$$\mathrm{S}_{ij} = \phi_{\text{img}}\left(v_i(\mathrm{x})\right)^\top \phi_{\text{text}}(p_j),$$

where x is the image, v_i one of the N transformations applied to it, and p_j one of the N text prompts. ϕ_{img} and ϕ_{test} are the CLIP visual and textual encoders, respectively.

This study enhances the evaluation framework by incorporating metrics and establishing relationships that directly correlate with observed errors. This facilitates a more detailed exploration of types and causes of errors, building upon the underlying insights provided in the original work.

3 Proposed Method

When reviewing the characteristics of machine generated content, a quantitative analysis can give insights into the efficiency and accuracy of a method in the form of statistical data. Yet for a creative task such as image generation, empirical metrics alone can not be relied upon to correctly illustrate output quality in every instance. Therefore a combination of quantitative and qualitative analysis can offer a more accurate representation [13].

The capacity of generative models to produce content at scale and with relative efficiency can be utilised in a quantitative evaluation step, where outputs are systematically preprocessed. Subsequent interpretation of the preprocessed data is carried out manually, introducing a qualitative view of the results. This approach is employed both for negative evaluation, to detect errors, as well as for positive evaluation of more complex generation tasks, to identify successful outputs. In particular, this constitutes a means for assessing the applicability of the colour channel permutation and non-cardinal rotation transformations. Figure 3 illustrates an overview of the evaluation pipeline.

With respect to computational efficiency, the feasibility of an alternating approach, rather than the current averaging method, is previously considered in the original work. The authors conclude that this method reduces output quality, especially for more elaborate generation tasks. This study deviates from the strict separation of the two approaches and endeavours to find a combination which balances sampling time and performance level.

Semi-Automated Evaluation Pipeline

Prior to conducting any evaluation, it is essential to define the criteria which distinguish successful illusions from unsuccessful ones. The following evaluation metrics build upon the metrics proposed in the paper by Geng et al.

Evaluation Metrics

Aligning prompts and transformations in the score matrix with their input order in the generator, a successful illusion is defined by large diagonal values in S and small off-diagonal values. Intuitively, this indicates that each transformation accurately depicts its corresponding target and suppresses other specified targets. The following metrics serve for a numerical interpretation of these characteristics:

The *alignment* score $\mathcal{A}$ measures the degree of association between corresponding transformation-prompt pairs:

$$\mathcal{A}(\mathrm{x}) = \min \operatorname{diag}(\mathrm{S}).$$

A high alignment implies that each target is well discernible in its respective transformation.

The *concealment* score $\mathcal{C}$ also accounts for off-diagonal values and is calculated as follows:

$$\mathcal{C}(\mathrm{x}) = \frac{1}{N} \operatorname{tr}\left(\operatorname{softmax}\left(\frac{S}{\tau}\right)\right)$$

where τ is the temperature parameter of CLIP. As a result, a higher concealment score indicates the suppression of undesirable targets when compared with the observation of the named target.

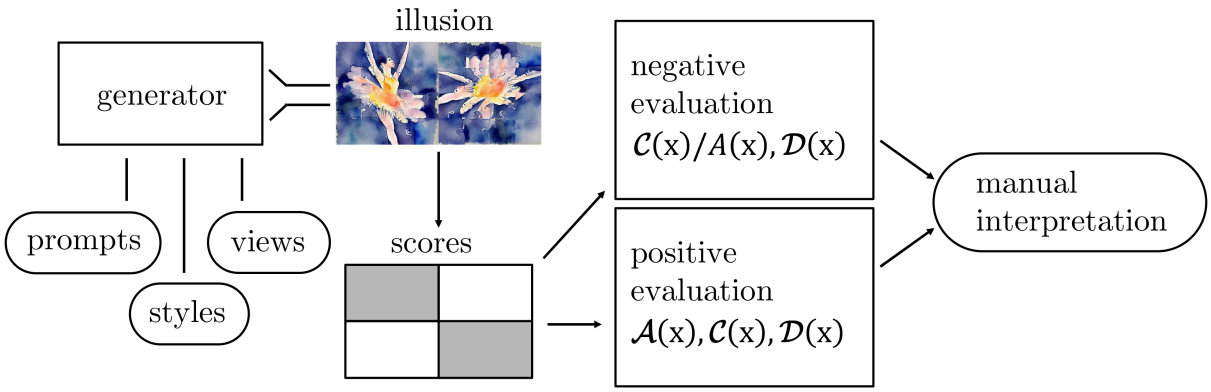


Fig. 3: Structure of the semi-automated evaluation pipeline. Input configurations comprising N prompts, N transformations, and a style are processed by the generative model to produce an illusion. The score matrix of the generated image is computed using CLIP. In the *negative evaluation* mode, configurations are filtered based on threshold values for $\mathcal{C}(\mathrm{x})/\mathcal{A}(\mathrm{x})$ and $\mathcal{D}(\mathrm{x})$ to identify errors. Conversely, in *positive evaluation* mode, configurations are filtered using thresholds for $\mathcal{A}(\mathrm{x})$, $\mathcal{C}(\mathrm{x})$, and $\mathcal{D}(\mathrm{x})$ to detect successful illusions. In both modes, automated preprocessing is followed by a manual analysis for further interpretation.

Finally, an additional metric is introduced within the scope of this study. The *dispersion* score $\mathcal{D}$ measures the variation in the most likely target across all transformations:

$$\mathcal{D}(\mathrm{x}) = \frac{1}{N} \operatorname{Var}\left(\operatorname{argmax}\left(\mathrm{S}_{i*}\right)\right)$$

where S_{i*} is the i-th row of S. Contrary to the concealment metric, this score disregards the index of the most likely target so long as the indices are dispersed. Simply put, a low dispersion score suggests the domination of one target across multiple transformations. This metric is in particular associated with the *dominant synthesis* error.

Negative Evaluation

The primary objective of this study is to identify the similarities between perception in diffusion models and in humans. Analyzing the failure cases of the model proposed by Geng et al., particularly the input configurations that result in unsuccessful illusions, offers valuable insights into patterns of discordance between machine-generated and human perception.

The first step of the negative evaluation pipeline consists of automated preprocessing. An *input configuration* comprised of N prompts, N transformations and an image style is passed into the model. The score matrix and $\mathcal{A}$, $\mathcal{C}$, and $\mathcal{D}$ scores are computed for the generated image. The configuration is flagged as erroneous if the values determined by $\mathcal{C}(\mathrm{x})/\mathcal{A}(\mathrm{x})$ and $\mathcal{D}(\mathrm{x})$ fall behind specified thresholds, which are determined through manual experimentation. This process is repeated for each combination of prompts, transformations and styles drawn from predefined lists until all possible configurations with $N \leq 4$ have been evaluated.

The components of a configuration can be viewed as features of the resulting illusion and accordingly, the above described procedure produces a characterisation of the conditions that lead to unsuccessful outcomes. A manual inspection of these results reveals patterns related to features or feature combinations that exhibit a heightened susceptibility to errors, as well as those that appear to mitigate them. These patterns offer insights into the characteristics of perception in diffusion models and serve as a basis for interpretation.

Positive Evaluation

Similarly to the error detection method, the evaluation pipeline can be used to identify successful illusions by adjusting the flagging condition. A result is considered successful when its alignment, concealment and dispersion score exceed predefined threshold values. A method for filtering successful configurations is particularly beneficial when analysing more complex inputs. For configurations incorporating the newly introduced colour channel permutations, non-cardinal rotations, or involving three or more transformations, the space of successful illusions is anticipated to be sparse. In such cases, the positive evaluation pipeline functions as a semi-automated mechanism for systematically exploring this constrained space.

Colour Channel Permutations

In the original work by Geng et al., transformations chiefly involve repositioning pixels to alternative locations within the spatial domain of the image. An exception is the *negation* transformation, in which pixel values are multiplied by -1 to produce the negative of the image. This transformation represents a type of illusion where diffusion models have the potential to surpass human artists in executing the same task, as visualising an image negative is inherently more challenging than a spatial transformation.

The rearrangement of the colour channels follows a similar intention by preserving pixel location and only altering their colour information. The validity of this transformation is conditioned on its linearity and statistical consistency: An image can be regarded as a three-dimensional tensor $\mathrm{x} \in \mathbb{R}^{H \times W \times 3}$ where H and W are height and width respectively and the final dimension defines the colour channels. A colour channel permutation can be expressed as

$$v_{\text{cc permute}}(\mathrm{x}) = \left(P\mathrm{x}^\top\right)^\top$$

where the inner transposition converts x to a $3 \times H \times W$ tensor and the outer transposition converts the result back to the original dimensions of $H \times W \times 3$. The permutation is performed by multiplying $\mathrm{x}^\top$ with an orthogonal permutation matrix $P \in \mathbb{R}^{3\times3}$ that rearranges the colour channels. Given that both transposition and multiplication operations are linear and P is orthogonal, it follows that the permutation of colour channel constitutes a valid transformation.

In practice, this type of transformation is implemented as a permutation to blue-red-green and one to green-blue-red. It is presumed that experimentation with two transformations is sufficient to ascertain the merit of the concept.

Non-Cardinal Rotation

Early experimentation with three transformations, specifically *identity*, *90°clockwise rotation* and *180°rotation* led to the observation that a difference of 90° between two rotations may not be sufficient to promote distinct targets. This motivated the implementation of a *120°rotation* and a *240°rotation* transformation, in the expectation that a greater degree of separation between the transformations facilitates more specific synthesis.

Because the diffusion model operates within a square layout, rotations of 120° or 240° disrupt this framework, resulting in invalid images. To address this limitation, images are first masked into a circular shape before applying the transformation.

The resulting transformation can be considered valid, as it bears similarities to the *45°inner circle rotation* described in the original study. This transformation is associated with an error known as *correlated noise*, suggesting that a similar error may arise with the newly implemented transformation.

Alternating Denoising

In its current configuration, the model estimates the noise for all transformations and targets at each inference step. It is established by Geng et al. that an alternating approach diminishes output quality, an observation that is based on experimentation with a prototype which does not fully capitalise on the potential of this concept [7]. While the observed reduction in quality concurs with the findings of the present study, the conclusion omits any effects that this concept has on sampling time. This evaluation seeks to either support or challenge this insight by incorporating the factor of complexity into the analysis.

Instead of performing all N estimates and disregarding all but one, the modified model selects the appropriate target and transformation for the current iteration and performs a singular noise estimation.

In order to find a balance between quality of the illusions and computational cost, a hybrid approach of alternating and averaging is implemented. A sampling process consisting of K inference steps in total is divided into a first stage of alternating comprising the first ℓ steps and a subsequent stage of averaging for the remaining $K-\ell$ steps. By adjusting ℓ and observing the mean sampling time and alignment score over numerous instances, an ideal point can be identified.

4 Experimental Results

The core focus of this study is to analyse optical illusions generated by diffusion models and draw conclusions about perception in artificial intelligence compared to human perception. Erroneous illusions highlight a disparities, arising when the model deems a result satisfactory, yet human evaluation judges it as flawed. Consequently, identifying patterns in configurations which lead to errors can provide insights into the reasons behind the observed differences.

Error Types

Geng et al. identify three types of errors arising from their model: *independent synthesis*, *noise shift*, and *correlated noise* [7]. The first is arguably the most insightful for this analysis, as it emerges independently of the model's implementation and hints at fundamental traits of diffusion models. Furthermore, experimentation uncovers a fourth type of error, termed *dominant synthesis*, which occurs in illusions that generate only one target. Figure 4 shows examples of each error.

Noise Shift

The noise shift error is linked to the *white balancing* transformation. This transformation does not to meet the criterion of statistical consistency and thus always results in an error. The authors hypothesise that this arises as a consequence of the model misinterpreting the scaled noise as signal.

Correlated Noise

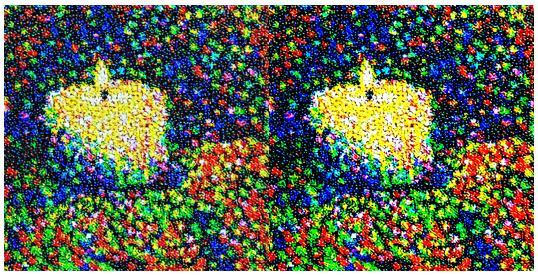
Most transformations specified in the spatial domain map pixels between integer coordinates. However, when this is not the case — such as with non-cardinal rotations — interpolation methods are required, which can introduce correlations in the noise. These correlations disrupt the inference process, preventing the successful synthesis of targets. This error type is similarly restricted to a particular class of transformations and can thus be attributed to the specifics of the model's implementation.

Independent Synthesis

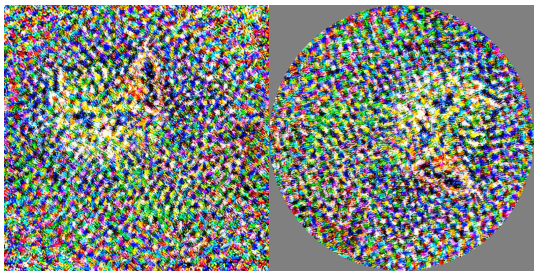
The phenomenon of independent synthesis warrants particular focus in this analysis. It describes a scenario where the generated image depicts the targets as individually synthesized elements, missing the integration into a unified illusion. This error is linked to a low ratio between $\mathcal{C}(\mathrm{x})$ and $\mathcal{A}(\mathrm{x})$.

A notable pattern observed in associated configurations involves the combination of targets that are animals, particularly those identifiable by their heads. Figure 4 provides an example of this phenomenon, featuring the specified targets *a rabbit* and *a duck*. Similar results arise from combinations involving *a cat*, *a dog*, *a pigeon*, and other animals. This type of error also occurs with human targets, such as *a young woman* and *an old woman* or *a prince* and *a princess*.

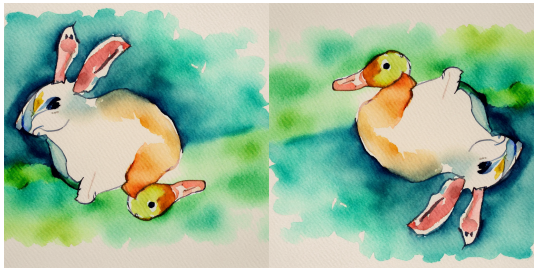
Another pattern that seems to promote independent synthesis is that of an object and an environment, provided the transformation preserves the majority of relative pixel positions. For a configuration containing targets such as *a kitchen* and *houseplants*, the model places the object (houseplants) inside the environment (kitchen).



(a) *Noise shift* error with targets *a candle* and *a wood fire*. Error is caused by the *white balancing* transformation which does not preserve Gaussian noise statistics.



(b) *Correlated noise* error with targets *a kitten* and *houseplants*. Error is caused by the *120° rotation* transformation that introduces correlation into the noise estimate through pixel value interpolation.



(c) *Independent synthesis* error with targets *a rabbit* and *a duck*. The combination of targets where both animals are identifiable by their heads leads to this error.

(d) *Dominant synthesis* error with targets *Elvis* and *a motorbike*. This error is likely a consequence of the semantic difference of the targets.

Fig. 4: Error types with associated configurations

Transformations that largely preserve the relative arrangement of pixels, such as rotations or flips, tend to be more prone to this type of error. In contrast, more disruptive transformations like jigsaw rearrangements or skewing exhibit a notable resilience.

An intriguing aspect of this error is that it cannot be strictly categorised as a flaw in the model itself. The model successfully accomplishes its task of synthesizing the specified targets but lacks an inherent understanding of what constitutes an illusion. Similar to humans, diffusion models appear to recognize objects primarily by shape [11], making it unsurprising that a partial result, such as a head, suffices for the guidance mechanism to identify the target within the image. The distinction from human perception, therefore, lies not in the ability to detect targets but in the capacity to comprehend the essence of an illusion.

These observations suggest that a creative task such as this requires a genuine grasp of what art truly entails. In this instance, it is ultimately humans who manipulate the diffusion model to produce illusions, and assign meaning to the resulting imagery.

Dominant Synthesis

A frequently encountered error during experimentation is referred to as *dominant synthesis*. This phenomenon occurs when one target becomes significantly more prominent than the others. A low dispersion score is indicative of this error type.

It should be noted that dominant synthesis is a frequent occurrence when automatically evaluating illusions. It appears to be the result of a semantic disparity between the targets. Combinations like *a penguin* and *a giraffe* can successfully create an illusion; however, when the prompts differ greatly in meaning — such as *Elvis* and *a motorbike* — one is typically favored during the diffusion process. It is therefore to be expected that this phenomenon occurs frequently, as most target combinations from the predefined list are essentially semantically dissimilar.

In contrast to independent synthesis, transformations that promote dominant synthesis act on a more complex level. Colour negation and pixel permutation appear particularly susceptible. This could be an indication that nuanced transformations such as these require thoughtful target selection to permit a successful illusion.

The high frequency with which this error type is encountered during automated generation implies that illusions cannot be produced autonomously or at a large scale. AI-generated art, though heavily reliant on machine assistance, still necessitates the involvement of a human operator to make critical design decisions.

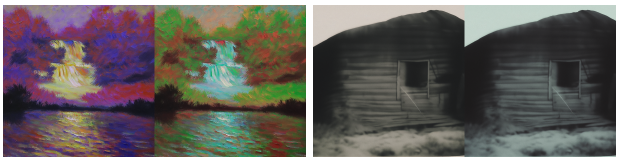
Three Transformations

Experiments involving three transformations and targets produce a limited number of favourable outcomes. For this task, the choice of transformations is crucial to the success of the illusion. When the transformations are contextually different, the model has greater potential at synthesising all three targets. Figure 5 shows a success case with three transformations. In many other instances, one or, at most, two targets predominate.

Colour Channel Permutations

For a transformation involving the rearrangement of color channels, selecting the appropriate targets is essential. An interesting occurrence with this type of transformation is the generation of a largely greyscale image when the targets

Fig. 5: Success case using three transformations. This illusion contains the targets *a girl*, *a young boy*, and *a dog* with transformations *identity*, *inner circle rotation*, and *jigsaw*. The inherent differences of the views result in the success of the illusion.



(a) Success case. Targets are *a waterfall* and *a sunset*

(b) Example of greyscale phenomenon observed. Targets are *a hut* and *a mountain*

Fig. 6: Examples of success and failure cases for the colour channel permutation transformation.

cannot be reconciled, as can be seen in Figure 6b.

Alternating Denoising

Figure 7 illustrates the effect that the length of the alternation period has on sampling time and alignment. As anticipated, longer alternation periods tend to decrease computational costs. In contrast to the findings of Geng et al., these results do not provide sufficient evidence to draw conclusions regarding the impact on output quality.

The effects on both metrics are minimal, with negligible impact observed. This is likely because the sampling stage represents only a small portion of the overall generation process, while the majority of the time is taken up by the upsampling stage, which is required in all cases.

5 Future Work

The evaluation of illusions mainly depends on numerical results, which may inherently reflect the same limitations as the diffusion model itself. An additional layer of abstraction could be achieved by implementing scoring or error detection through a large multimodal model, either by leveraging a pre-trained model with tailored prompts or by training the model specifically for this task. This approach would require careful consideration of optimisation techniques, as the extended inference time of large multimodal models currently renders their use impractical.

With respect to the sampling time of the diffusion model, the alternating approach yielded unfavourable results. An interesting path for further research is the exploration of different optimisation methods.

Finally, although the implementation of non-cardinal rotation transformations directly led to issues, the underlying concept remains worthy of exploration. Developing an approach that eliminates the need for interpolation could enable the successful application of this transformation.

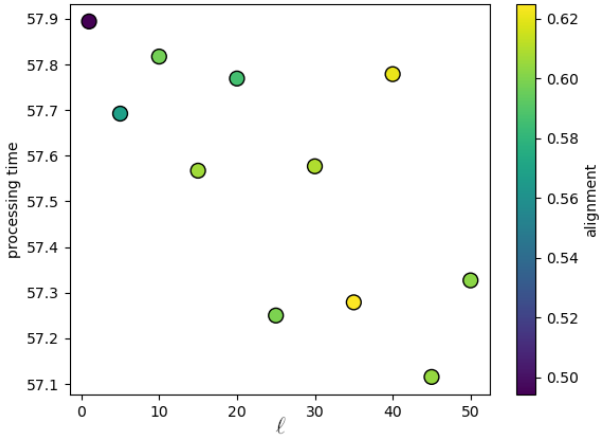


Fig. 7: Effect of length of alternating stage on processing time in seconds and alignment score.

6 Conclusion

Image generation with artificial intelligence has experienced rapid advancement in recent times and diffusion models in particular show great potential. When *perception* is presumed to refer to surface level tasks such as object detection, the findings of this study align with previous research — there appear notable similarities between diffusion models and humans [7, 11, 12].

The fundamental difference then lies not in perception but in understanding the purpose of an artwork. Diffusion models can produce images at large scales, but human artists comprehend the meaning of an image on an abstract level of which an artificial intelligence model is not capable.

Future advancements in this domain may enable generative models to more effectively emulate aspects of what is often considered sentience, potentially expanding into the creative realm of artistry. This development would undoubtedly spark ethical debates, raising questions about whether machine-generated art can truly be deemed creative and whether the autonomy of such processes might intrude upon the uniquely human joy of creative expression — a domain that some might argue should remain untouched.

A more optimistic prognosis is that generative models will follow an evolution similar to that of the photo camera, in that they become a tool for a novel creative discipline, simultaneously sparking diverging movements in traditional art forms. A first indication of this development is work by artists such as MrUgleh who use generative models as an instrument for creative expression [15].

Acknowledgments

The author would like to thank David Lindell, Ph.D. and the CSC2529 teaching team at the University of Toronto for their support and enthusiasm. Special thanks to Aviad Levis, Ph.D. for continuous assistance and creative input.

References

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021.

[2] K. N. Sai, U. Wable, A. Singh, K. N. V. S. S, H. Gonuguntla, and C. Jain, "Harnessing multimodal ai for creative design: Performance evaluation of stable diffusion and dall-e 3 in fashion apparel and typography," in *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, 2024, pp. 1–6.

[3] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," 2022.

[4] J. Oppenlaender, "The creativity of text-to-image generation," in *Proceedings of the 25th International Academic Mindtrek Conference*, ser. Academic Mindtrek '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 192–202.

[5] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 15 903–15 935.

[6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[7] D. Geng, I. Park, and A. Owens, "Visual anagrams: Generating multi-view optical illusions with diffusion models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 674–10 685.

[9] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 78 723–78 747.

[10] A. Gómez-Villa, A. Martín, J. Vazquez-Corral, and M. Bertalmío, "Convolutional neural networks deceived by visual illusions," *CoRR*, vol. abs/1811.10565, 2018.

[11] P. Jaini, K. Clark, and R. Geirhos, "Intriguing properties of generative classifiers," 2024.

[12] J. Ngo, S. Sankaranarayanan, and P. Isola, "Is CLIP fooled by optical illusions?" 2023.

[13] A. Borji, "Qualitative failures of image generation models and their application in detecting deepfakes," *Image and Vision Computing*, vol. 137, p. 104771, 2023.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[15] Ugleh. (2023) Spiral town - different approach to qr monster. [Online]. Available: https://www.merriam-webster.com/thesaurus/leverage