Seeing The Words: Evaluating AI-generated Biblical Art

Hidde Makimei[1][0009-0004-7184-7943], Shuai Wang[1][0000-0002-1261-9930], and Willem van Peursen[3][0000-0002-3142-5752]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
hidde.n.g.makimei@student.vu.nl| shuai.wang@vu.nl

[2] Eep Talstra Centre of Bible and Computer (ETCBC), Faculty of Religion and Theology, Texts and Traditions, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
w.t.van.peursen@vu.nl

Abstract. The past years witnessed a significant amount of Artificial Intelligence (AI) tools that can generate images from texts. This triggers the discussion of whether AI can generate accurate images using text from the Bible with respect to the corresponding biblical contexts and backgrounds. Despite some existing attempts at a small scale, little work has been done to systematically evaluate these generated images. In this work, we provide a large dataset of over 7K images using biblical text as prompts. These images were assessed under multiple neural network-based tools. We provide an evaluation of the accuracy and some analysis from the perspective of religion and aesthetics. Finally, we discuss the use of the generated images and reflect the performance of the AI generators.

Keywords: Biblical Art · Generative AI · Computational creativity · image processing.

1 Introduction and Related Work

Despite the fact that more and more Artificial Intelligence (AI) tools can generate art, the generated images have been argued to lack human attributes such as creativity, originality, subjectivity, emotional depth, context, cultural significance, intention, and conceptualisation [5, 6, 13]. The biblical text has served as a wellspring of inspiration for human creativity across various domains. Its stories, metaphors, ethical teachings, and representations of divine creation have guided and fueled the imagination of artists. Recently, there has been some primitive work about using AI-generated biblical art. The BiblePics App[3] takes advantage of AI-generated images and provides visualized scenes in the bible. As far as the authors know, the largest collection of generated art [1] is hosted on the OpenBible website[4]. There are 1,128 images generated using DALL·E

[3] https://biblepics.co/

[4] https://www.openbible.info/labs/ai-bible-art/

2 including contributions from communities. Their corresponding prompts were not given, which makes the assessment of context and objects impossible. Figure 1 shows that the AI-generated picture can be quite accurate and similar to the common imagination of the scene implied in the text prompt. Figure 2 shows how the generated images may also include surprising elements such as sky-scrapers in the background of the Last Supper. However, such an approach to biblical stories, filling in the background with contemporary elements is also well-known in human art. An example are the medieval setting (clothes, castle in the background), in Figure 3.

As far as the authors are aware, none of the existing work includes an analysis of how accurately these generated images correspond to the text, nor about their aesthetics. This raises the need for a systematic assessment of images produced in this approach. A comparison of AI-generated images with well-known paintings by artists on the same topic can help understand the confounding differences between AI and humans, as well as analyze the bias of generators and guide the development of future AI-based tools. This comparison could also guide the selection of relevant images and ease manual evaluation.

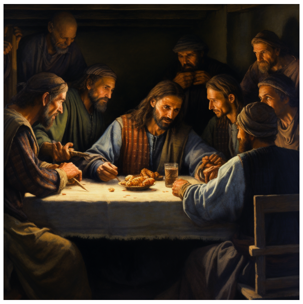

Fig. 1: An image about the Last Supper generated by Midjourney in our VDD dataset

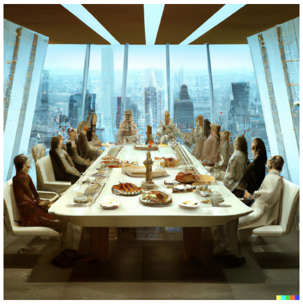

Fig. 2: An image about the Last Supper generated by Dall E, provided by OpenBible

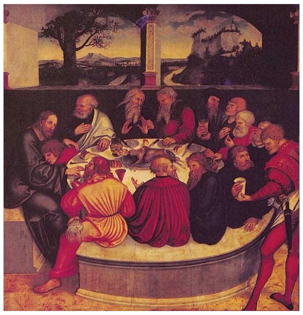

Fig. 3: The Last Supper painted by Lucas Cranach the Elder in 1547

The research questions of this paper are the following. **RQ1:** How can we systematically generate biblical images using text-to-image generators? **RQ2:** How can we evaluate the biblical images generated? For this question, we perform the evaluation in three aspects using the subquestions: **SRQ2A:** What is the accuracy of persons and objects in the generated images regarding their biblical context? **SRQ2B:** How can we compare the sentimental values of the generated images? **SRQ2C:** What features can we observe for the generated images regarding religion and aesthetics?

This paper presents the *Visio Divina Dataset* (VDD in short)[5], a large open dataset of images generated using various AI generators. The dataset consists of 7,116 images from 9 text-to-image generators. Selected images are included in an online Virtual Reality (VR) exhibition.[6]

We make the first attempt to construct a workflow and incorporate automated evaluation of AI-generated images against well-known paintings by artists referring to the same biblical text. The paper presents the results of (manually or automatically) evaluation of several features: accuracy evaluation, sentimental analysis, religious analysis, and aesthetic analysis.

This paper is organized as follows: Section 2 provides details on the selection of prompts and the generation of images. Section 3 includes the pipeline and evaluation metrics. Section 4 presents the evaluation results. In Section 5, we discuss the findings and limitations of the approach. Section 6 presents the conclusion and future work.

2 Data

2.1 Prompt

To unify the input of text-to-image generation, we select some representative biblical themes that have been studied by artists with a rich amount of masterpieces. More specifically, we take five different passages as prompts. The prompt selected correspond to these themes: 0) Adam and Eve's Expulsion of Paradise (Genesis 4:23-24) 1) The Tower of Babel (Genesis 11:1-9) 2) Binding of Isaac (Genesis 22:9-14) 3) The Last Supper (Mark 14:12-25), 4) Moses Found Exodus 2:5-9). These prompts correspond to the aspects to be assessed with details in Section 3.

2.2 Image generation

Since the existing work shows no systematic generation of biblical art, for a fair assessment, it is essential to provide a dataset using the same input under the same settings for all the generators. We select some state-of-the-art generators including Dall-E 2, Midjourney as well as seven different versions of Stable Diffusion. For the best performance, we used the commercial version of Dall-E 2 [7] and Midjourney [8]. For Stable Diffusion, we used some popular open-source

[5] Visio Divina is a practice of the form of divine seeing by prayerfully inviting God to speak while looking at an image. For this submission, we provide an anonymous repository with data, code, and supplementary material on Figshare: `https://figshare.com/s/829b5d07b524690cb5a2`. Links to public repositories will be included after the paper gets accepted.

[6] `https://shuai.ai/art/seeing`

[7] `https://openai.com/dall-e-2`

[8] `https://www.Midjourney.com/home/`

tools: SG161222 (SG in short)[9], runwayml (RW)[10], CompVis (CV)[11], stabilityai (SAI)[12], prompthero (PH)[13], nitrosocke (NS)[14], and dreamlike-art (DA)[15]. Since Midjourney lacks an API, we customized a bot that takes over the computer and interacts with the Midjourney bot for the automatic collection of generated images. For Dall-E 2, we used its API. For all variants of Stable Diffusion, we generated the images on the Google Colab cloud server that uses the A100 GPU. All the generators were accessed in the week of 19th of June, 2023. All the images are associated with a unique code for easy reference.

The images were produced through an automated process where the prompts were fed repeatedly into the generators with a summary in Table 1. For DALL-E 2, the size of the prompt exceeded the character limit. Thus, prompts 1 and 3 were reduced by using NLTK Library[16] with stopping words and punctuation removed.

Table 1: A summary of AI generators and their generated images

	Dall E	Midjourney	Stable Diffusion							Sum (VDD)
			RW	CV	SAI	PH	SG	NS	DA	
Version	V1 beta	V5.1	V1.5	V1.4	V2.1	V1.1	V1.4	V1.1	V2.0	
#images	500	616	1K	1K	1K	1K	1K	500	500	7,116

2.3 Artwork

The biblical artwork chosen to compare to the AI-generated images are paintings from the Renaissance and Baroque periods. Choosing a time period narrows down the sample group of biblical art to more similar like-minded artists. The Renaissance shows the emergence of a naturalistic style (compare, e.g., the interest that painters developed in anatomy, proportions and perspective). This was further developed in the Baroque, which is well-known for its use of contrast, movement, exuberant detail, deep colour, grandeur, and surprise to achieve a sense of awe, in other words, to express sentiment. These features render paintings from these style periods good candidates for automatic analysis (e.g., object or sentiment recognition).

Moreover, some of these paintings, like Leonardo da Vinci's Last Supper or Pieter Bruegel's Tower of Babel belong to the most famous works of art

[9] https://huggingface.co/SG161222/Realistic_Vision_V1.4
[10] https://huggingface.co/runwayml/stable-diffusion-v1-5
[11] https://huggingface.co/CompVis/stable-diffusion-v1-4
[12] https://huggingface.co/stabilityai/stable-diffusion-2-1
[13] https://huggingface.co/prompthero/openjourney-v4
[14] https://huggingface.co/nitrosocke/Ghibli-Diffusion
[15] https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0
[16] https://www.nltk.org/

history and have shaped the Western imagination of these scenes (which, as we shall see, to some extent also affects the AI-generated images). Choosing a time period narrows down the sample group of biblical art to more similar like-minded artists. The selection of the paintings of was done by considering the visibility of characters and accessibility to give the art a fair possibility for the machine learning models to evaluate it.

Table 2: A summary of the authors and the years of the paintings chosen for this paper

Prompt 0	**Prompt 1**	**Prompt 2**	**Prompt 3**	**Prompt 4**
Michelangelo (1512)	Bartolomeo Cavarozzi (1598)	Lucas van Valckenborch (1594)	Juan de Juanes (1560)	Lucas van Valckenborch (1635)
Jan Bruehel (1624)	Lucas Gassel (1539)	Pieter Breugel (1563)	Peter Paul Rubens (1632)	Toussaint Gelton (1645)
Benjamin west(1760)	Carvaggio(1598)	Grimmer(1604)	Il Tintoretto(1592)	Jan Kosten(1650)
Izaak van Oosten(1628)	Titiaan(1542)	Hendrick van Cleve(1570)	Hans Holbein de Jonge(1527)	Paolo Veronse(1570)
Cornelis van Poelenburg(1652)	Rembrandt(1635)	Frederik van Valckenborch(1600)	Leonardo da Vinci(1495)	Bartholomeus Breenbergh(1622)

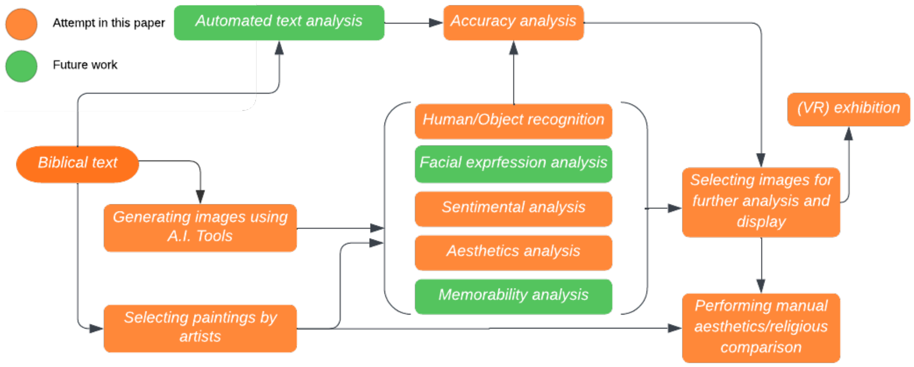
Attempt in this paper
Future work
Automated text analysis
Accuracy analysis
Human/Object recognition
Facial exprfession analysis
Sentimental analysis
Aesthetics analysis
Memorability analysis
(VR) exhibition
Biblical text
Generating images using A.I. Tools
Selecting paintings by artists
Selecting images for further analysis and display
Performing manual aesthetics/religious comparison

Fig. 4: Workflow of image generation and evaluation against paintings

3 Methodology

Figure 4 is workflow that visualizes the steps taken in the study. For the automated analysis, we focus on two aspects: people and sentiment. As for the people in the generated images, we take advantage of state-of-the-art neural network models for the evaluation of the number of people as well as their ages and gender. In addition, we discuss the aesthetics by comparing it against selected masterpieces. Table 3 provides a summary of the CNN (Convolutional

Neural Network) models used to obtain values. Excluded from this study are the weather, the style of clothes, objects beyond human beings (e.g. angels and ghosts), the facial expression of the people, and other aspects that are either difficult to evaluate or not directly related to the biblical context. Although there are models for predicting the genre [2] and the style [14], they are beyond the biblical context in this study. Small objects such as knife, apple are mentioned in the prompts and could be added to the workflow but will be left for future work.

Table 3: Models used and their training datasets

	Aspect	**(Core) Model(s)**	**Training Dataset**
Part 1: Human Analysis	Number of people	Detectron2 (Mask R-CNN and ResNet-50)	COCO, Cityscape
	Age	LeNet-5	ImageNet
	Gender	LeNet-5	ImageNet
Part 2: Sentimental Analysis	Sentimental value	AlexNet	Twitter Dataset

3.1 Part 1: Human Analysis

To answer our research question SRQ2A, for the analysis of human beings in the images and paintings, we focus on three aspects: the number of humans as well as their age and gender.

Human Recognition with Detectron2 Detectron2 [15] is a Mask R-CNN (Region-based Convolutional Neural Networks) with both ResNet50 [8] and FPN (Feature Pyramid Network) [11] as its backbone. It uses Mask R-CNN and extends the Faster R-CNN model by masking in order to achieve pixel-wise segmentation. The Mask R-CNN includes four layers of 3x3 convultion applied to a 14x14 input feature map, whose output passes through a deconvelution layer which gets transformed using a 2x2 kernel and ends with a 1x1 convolution network that predicts the mask logits. This model is used for mapping the segmentation and is trained on the COCO dataset [12] with 8 categories and the Cityscape dataset [7] and predicted using the backbone model. We only identify the label corresponding to human to be found for each given image and use those with a confidence score of 0.8 or above. The outputs are some bounding boxes for each person detected, which are used to count the number of person detected in the images in this study. [performance]

Age and Gender Estimation For age and gender estimation, a custom CNN [10] was developed by Gil Levi and Tal Hassner based on LeNet-5, whose main architecture consists of three convolutional layers and two fully connected layers. Each layer of the CNN is followed by ReLu and normalization before being passed on to the next. Finally, the fully connected layers are mapped to the final phase that can classify the age and gender respectively. For this model, the Imagenet dataset was used for training. The network produces an age prediction in the form of a range with a minimum and a maximum. In this paper, the detected human in bounding boxes Human Recognition with Detectron2 the estimated age is taken as the average of the minimum and the maximum. For this work, we take the predicted gender: male or female. Non-binary cases are beyond the scope of this work. [Todo: talk about the accuracy of this model]

3.2 Part 2: Sentimental Classification

To answer the research question SRQ2B, for sentimental recognition, we use a model introduced by Victor Campos et al. [3] Using an AlexNet-styled network [9] composed of five convolutional layers and three fully-connected layers. The model passes the pixel value through the CNN to obtain an overall sentimental value of the image. It takes the Twitter dataset as training data [16]. One observation is that it tends to map brighter pixels to more positive sentiment. The resulting sentimental value is in an interval between 0-1 (1 for positive). The model can be altered into a fully convolutional network with no additional training need. This produces kernel 8x8 predictions maps of the image giving 64 patches of the image with its own sentimental value. Given that the resulting sentimental values would differ if the two networks differ, we evaluate the result on two different settings.

3.3 Metrics

Next, we introduce some metrics to unify the ouput of evaluation models. To do this, we first transform the output from neural network models into a number in the interval of [0,1]. The results from each of the following measures assessing different aspects are then integrated. Next, we provide the details of these measures.[17]

Number of people Recall for each generated image, we obtain the number of people detected, denoted n as described in Section 3.1. We compute its "distance" to each selected human art by computing the difference with the number of people in it. We then divide its absolute number by the maximum number of people detected N among both the generated images and human art. This transforms the difference into an interval form with a score between 0-1. We then take the mean for each generator, which is an average difference in the generated artwork against the selected human art.

Gender For female and male, we compute two numbers respectively. Take the number of females for example, for each generated image, we compute the difference in the number of detected females regarding each painting. The average of the absolute value was taken and divided by the maximum number of females

[17] More details and intermediate results are given in the supplementary material.

among all the generated images and paintings to unify this number to the interval of 0-1. For each generator, the resulting average and standard deviation for all its generated images for each prompt show the difference in the number of females generated as well as its diversity. Same for that of male.

Age Since numbers of age are categorized in ten groups, we calculate the differences in each group between a given human artwork and a generated image. We then divide its absolute value by the maximum number of people detected. The rest is similar as that of the number of people described above.

Sentimental values and scores We compute two scores for sentimental analysis. For each generated image, we obtain the *sentimental value* as described in Section 3.2. Similar to the calculation steps as described above, the sentimental value for each generator is the average of the generated images about the absolute difference regarding each human art for each prompt. For the second score, we compute each patch (a section of the image at the same location) between the generated image and human art. The average of all the patches is then the score of this comparison. For each prompt, the overall *sentimental score* for a generator is the average of the difference of the sentimental value of each pair of generated images and paintings.

Finally, as a proof-of-concept, we take the overall score simply as the average of all the above-mentioned scores for all the aspects assessed. The lower the scores are, the more the generated images are like the selected artwork.

4 Evaluation

Next, we provide evaluation details and compare the scores under different settings as introduced in Section 3. Since the evaluation of religious aspects and aesthetics cannot be done fully automatically, they are manually assessed and included in Section 5.

Number of people, Gender and Age Figure 5 is the proportion of recognized humans in the generated images for each generator regarding prompt 3 (the Last Supper).[18] The green bars indicate that Midjourney can generate images with more people while those by Dall E (shown in red) are very unlikely to have humans recognized. We also present two variants of Stable Diffusion to show that there can be more humans in the generated images (e.g. RW) as well as fewer (e.g. SG). While more people can be found in the human artwork. This shows that Dall E lacks an understanding of the biblical context and the characters while that of Midjourney and Stable Diffusion could have used. Moreover, we noticed that, oftentimes, it is the case for Midjourney that a middle-aged man (representing Jesus) is around the center of the image with a few others surrounding him (e.g. Figure 1).

Table 4 shows in detail the standard deviation (STD) and mean of the number of males and females as well as the total number of people detected for

[18] Similar evaluation for all the prompts and their details can be found in the supplementary material.

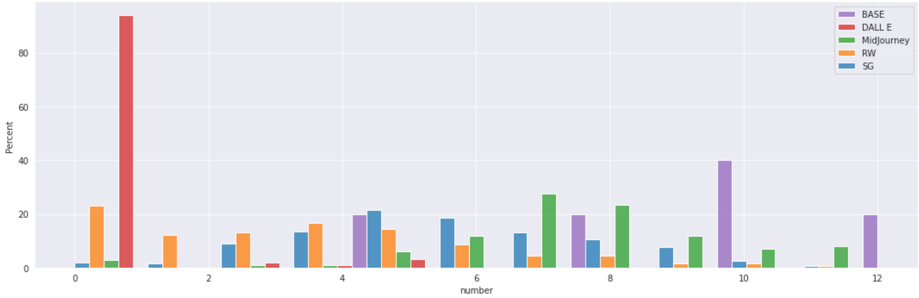

Fig. 5: The proportion (in percentage) of the number of people recognized in all generated images by selected generator for prompt 3 (the Last Supper)

prompts 3 and 4 in which males and females are dominant respectively so that the recognition of gender for Midjourney and Stable Diffusion has a similar mean to the human art. The STD for male recognition is on average much higher across all generators for prompt 3. This is a weakness recognized in the gender classification since some long-haired male characters are classified as females. Age followed a similar pattern the report shows the distribution of age being similar in the human artwork and Stable Diffusion and Midjourney with DALL E lacking for this aspect as well. Finally, the last row in Table 4 shows MidJourney having the highest mean in the number of people recognized and mean for males recognized which reciprocates the high number recognized by the human art showing similarity in human incorporation in the artwork. Stable Diffusion has higher STDs, indicating that the most diverse images are generated. While DALL E performs poorly as seen with its low score in STD and mean.

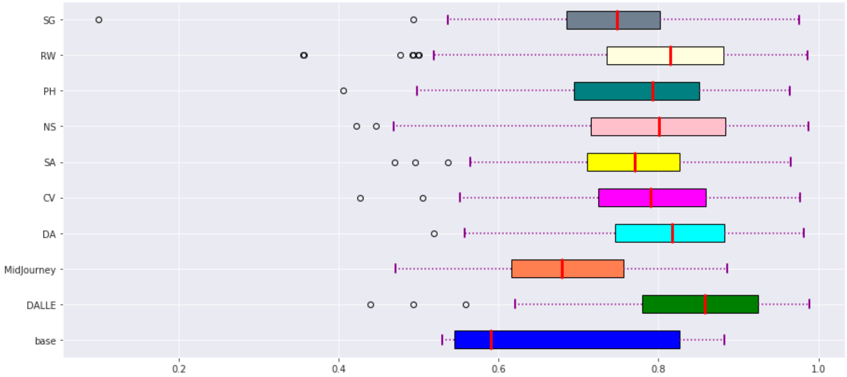

Fig. 6: The sentimental value for prompt 1 by neural network models (0 represents negative and 1 for positive)

Table 4: Comparing the mean and standard deviation for the assessment of gender and number of people as well as the overall accuracy for the generated images. M-STD standards for the standard deviation of the number of males detected in the images. Similarly, F-STD is for that of females. N-Mean refers to the average number of people detected in the image. N-STD is its standard deviation. The highest values are in bold font and the lowest are underlined.

		Base	Midjourney	Dall E 2	CV	PH	SAI	DA	NS	SG	RW
Prompt 3	M-STD	2.3021	2.1887	0.3258	2.0220	2.0220	**2.2826**	2.1082	1.5973	1.8842	2.0827
	M-Mean	7.6000	**6.9607**	0.0700	2.1800	2.1800	3.3400	2.1400	1.4650	4.5850	2.6550
	F-STD	0.7071	1.0151	0.2428	0.8954	0.8954	**1.0677**	0.6590	0.8896	0.9766	1.0296
	F-Mean	2.0000	**1.1372**	0.0400	0.8100	0.8100	0.8400	0.5000	0.7500	1.0300	0.9850
	N-STD	2.7928	2.3902	0.4691	2.4308	2.4308	**2.7248**	2.1997	1.8991	2.1887	2.5243
	N-Mean	9.6000	**8.0980**	0.1100	2.9900	2.9900	3.3400	2.6400	2.2150	5.6150	3.6400
Prompt 4	M-STD	2.1213	0.8643	0.2777	0.7166	0.6887	**0.7499**	0.5222	0.3896	0.7421	0.6244
	M-Mean	2.0000	0.3437	0.0600	0.5400	0.4800	0.5200	0.3000	0.1700	**0.5450**	0.4550
	F-STD	0.8944	0.9636	0.5773	0.9519	0.9101	1.1072	0.7177	0.8084	0.8035	**0.9664**
	F-Mean	1.4000	**2.4765**	0.5000	1.2700	2.0000	1.5100	1.5000	0.6400	1.7600	1.2750
	N-STD	2.8809	0.9835	0.6407	1.1164	0.8466	**1.2233**	0.6030	0.8704	0.5599	1.0783
	N-Mean	3.4000	**2.8203**	0.5600	1.8100	2.4800	2.0300	1.8000	0.8100	2.3050	1.7300
Overall average across all prompts	M-STD	1.3400	1.3585	0.2821	1.1991	1.3266	**1.3930**	1.0729	1.1701	1.2231	1.1581
	M-Mean	2.4800	**2.6317**	0.0600	1.1020	1.2780	1.4010	1.2080	0.7520	1.7970	1.0350
	F-STD	1.1602	0.9278	0.2853	0.8689	0.9247	1.0966	0.6533	0.8099	**1.1300**	0.6558
	F-Mean	1.4000	1.0078	0.1400	0.6800	0.9480	0.8670	0.7220	0.5130	**1.0250**	0.6440
	N-STD	2.0705	1.5313	0.4484	1.5929	0.8466	**1.9731**	1.1896	1.3275	1.7029	1.5115
	N-Mean	3.8800	**3.7759**	0.2000	1.7820	2.2260	2.1000	1.9300	1.2650	2.8220	1.6790

Sentimental analysis As shown in Figure 6, the base paintings seem to have a more neutral sentimental mean value than those of the generators. Here, Midjourney shows the most similar score as the base paintings. When compared against human art regarding the sentimental score (the difference between the sentimental value in human artwork and generated images). Table 5 shows the sentimental score for prompt 1 and 4, which indicates that those by Midjourney is the most similar to the human artwork. In contrast, the sentimental score by DALL E differs most from the base paintings.

Table 5: Comparing the average sentimental scores for prompts 1 and 4

	Midjourney	Dall-E 2	CV	PH	SAI	DA	NS	SG	RW
Prompt 1	0.1522	0.1908	0.1777	0.1718	0.1658	0.1751	0.1834	0.1645	0.1897
Prompt 4	0.1491	0.1739	0.1588	0.1524	0.1512	0.1629	0.1687	0.1496	0.1516

Overall Scores The overall results per prompt can be seen in Table 6. The score shows the average of all the calculated scores, giving us an indication of the overall difference between the human artwork against their respective AI counterparts.[19] It shows that Midjourney is the most similar to the selected human artwork. Midjourney scores the best in the sentimental score for all prompts.

[19] For details analysis see supplementary analysis

In addition, the score for age, gender and number of people is also one of the best. DALL E performs the worst across every metric scoring section. With it lacking the capabilities to produce recognizable human characters that the CNN model is able to detect and in turn not being able to score human characteristics. On the contrary, Stable Diffusion similarity depends on the prompt with some prompts producing more similar paintings to the human artwork. Its variations show a slight difference in score per prompt input. The last row in Table 6 is for the overall scores. It shows the average score tallied up from all prompts. The smaller the score, the better the performance. As we can see, Midjourney gives the best overall score. Its performance is significantly better than that of various versions of Stable Diffusion and outperforms Dall-E.

Table 6: Comparing the overall score of different generators for each prompt

	Midjourney	Dall-E 2	CV	PH	SAI	DA	NS	SG	RW
Prompt 0	0.1148	0.1852	0.1208	0.1237	0.1261	0.1275	0.1285	0.1417	0.1508
Prompt 1	0.1196	0.1384	0.1340	0.1309	0.1347	0.1302	0.1358	0.1322	0.1384
Prompt 2	0.1219	0.1333	0.1219	0.1407	0.1223	0.1515	0.1245	0.1364	0.1216
Prompt 3	0.1286	0.2490	0.1926	0.1662	0.1743	0.2019	0.2087	0.1508	0.1788
Prompt 4	0.1448	0.1883	0.1690	0.1613	0.1657	0.1698	0.1840	0.1597	0.1665
Overall	**0.1279**	0.1788	0.1477	0.1446	0.1446	0.1562	0.1563	0.1460	0.1512

5 Discussion

In this work, our selected paintings by artists prioritized realism or naturalism, which introduces a bias. Based on our metric that integrates all aspects under assessment, we concluded that Midjourney performs the best overall when compared to the selected paintings. There could be several reasons for this. Images generated by Midjourney have more details, which improves the accuracy of the recognition of objects, e.g. faces, especially those partially hidden or under the shadow. This corresponds with the interests in the Renaissance and Baroque periods (cf. section 2.3). The Midjourney images could be easily adopted with little modification as illustrations for biblical blogs, books, etc. Stable Diffusion is different in that some images exhibit some level of abstraction.

More details do not always imply better accuracy. We noticed, for example, that *the cross* can be shown in the same generated image as Jesus. As a symbol of the Christian faith, this is understandable, but displaying a cross in a scene preceding Jesus' death is highly anachronistic. Details can also imply challenges at the detail level. Thus, despite that the version of Midjourney has been fine-tuned for the generation of hands, none of the generators can generate perfect hands. We often observe polydactyly (one or multiple supernumerary fingers).

Another reason for the differences could be due to the style of generated images. Those by Midjourney exhibit the art style of the Renaissance period, which could be the result of the inclusion of Renaissance art in its training data. The Dall-E 2 images can vary significantly in style, which could reveal some degrees of creativity. This touches upon the general questions as to how we should define creativity in relation to accuracy. Do we assess the Tower of Babel depicted as a skyscraper or skyscrapers in the background of the Last Supper scene (Figure 2 as "inaccurate" or as "creative", or both?

In this work, we do not combine our work with the analysis of text. Thus, the semantic correspondences were analyzed manually (but could be done with automation in the future). We noticed that the differences are significant: DALL-E seems to be unable to make sense of some text prompts and their context. Midjourney and Stable Diffusion perform better, but differently. The training of Stable Diffusion seems to rely on traditional paintings of the biblical scene, whereas Midjourney picks up the building activity in a relatively naturalistic way (like cartoons, illustrations in Children's Bible). Some semantic aspects of the text prompts apparently posed challenges. Thus "The twelve" in the text as reference to the twelve disciples is in most cases not picked up. The text prompt contains some concrete objects (cup, bread), but also much conversation. This may have evoked the confusion that is especially visible in the DALL-E Images. The context often appears in Dall-E as words overlaying on some background images. In most cases, these texts are hard to recognize. Only when language and visual communication play an important role (as in the Tower of Babel story), we may see a link between the text prompt and the text as part of the generated image.

6 Conclusion and Future Work

In this paper, we take an interdisciplinary approach for the study of AI-generated biblical art. We performed a systematic evaluation of the images generated using biblical text. For RQ1, we selected biblical text as prompts and generated a large dataset with over 7K images. Moreover, we chose five paintings for each prompt as references for evaluation. The RQ2 was tackled with the help of different neural network-based image assessment models. We proposed metrics for the assessment of accuracy. Our analysis answered SRQ2A. As for SRQ2B, we employed two models to obtain the sentimental values. For SRQ3C, we provide an analysis regarding religion and aesthetics. **Overall, Midjourney generates the best images when assessed using our metrics and selected paintings.** Among all the variants of Stable Diffusion, prompthero and stabilityai give the best images. In contrast, Dalle E is the worst generator. Finally, we discussed the limits of our approach, reflected on the evaluation results, and discussed the features of the AI generators.

There are many issues that require further investigation. There are objects other than humans in the selected biblical text including altar, wood, knife, as well as spiritual beings. Midjourney and Stable Diffusion perform relatively well on generating such objects based on our manual assessment. The recognition of typical scenes such as the Last Supper or the Tower of Babel as well as the way in which we could manipulate and improve the prompts, deserves further study. Some prompts had truncated text because of the maximum size of the prompt.

So we could further assess images produced based on different truncated versions. We noticed that in some images, due to the long hair, some males were recognized as females. This shows that the models' performance needs to be evaluated on the generated images. Otherwise, its errors can have a negative impact on the evaluation result. Moreover, the workflow could be extended to incorporate additional machine-learning and image-processing models that classify landscape, facial emotion, and weather, and study how the generated images vary in context, accuracy, art style, theme, and other interpretative features. In addition, some deep learning models can be used to evaluate memorability [4], which could be integrated into the current evaluation scheme.

Further analytical results of biblical text could be achieved by comparing different versions of the Bible and evaluating the images produced regarding some traditional versions (e.g. the King James Version) and modern translations (e.g. Good News Bible). It remains to be studied how we could use painting beyond the Renaissance and Baroque artworks to evaluate the performance. Finally, the generated images as well as the assessment results could be used for future research and benchmarking the performance of generators on topics in the intersection of art, theology, and computer science.

Finally, although our dataset is published as an open source, finding a useful image with specific features can be hard due to its large size. We plan to create an indexing and searching platform to make it possible to retrieve images with certain features (e.g. six males, two or three females, and a given sentimental value) according to the scores computed as described in Section 4. This platform could benefit users, especially artists, to easily retrieve images of interest.

References

1. Openbible: AI-generated bible art (2023), https://www.openbible.info/labs/ai-bible-art/
2. Agarwal, S., Karnick, H., Pant, N., Patel, U.: Genre and style based painting classification. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 588–594. IEEE (2015)
3. Campos, V., Jou, B., i Nieto, X.G.: From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. Image and Vision Computing **65**, 15–22 (2017). https://doi.org/https://doi.org/10.1016/j.imavis.2017.01.011, https://www.sciencedirect.com/science/article/pii/S0262885617300355, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing
4. Cetinic, E., Lipic, T., Grgic, S.: A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access **7**, 73694–73710 (2019)
5. Chatterjee, A.: Art in an age of Artificial Intelligence. Frontiers in Psychology **13**, 1024449 (2022)
6. Cheng, M.: The creativity of Artificial Intelligence in Art. In: Proceedings. vol. 81, p. 110. MDPI (2022)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
10. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops (June 2015), `https://osnathassner.github.io/talhassner/projects/cnn_agegender`
11. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
13. Liu, B.: Arguments for the rise of Artificial Intelligence Art: Does AI art have creativity, motivation, self-awareness and emotion? Arte **Avance en línea**, 1–11 (04 2023). https://doi.org/10.5209/aris.83808
14. Shamir, L., Tarakhovsky, J.A.: Computer analysis of art. Journal on Computing and Cultural Heritage (JOCCH) **5**(2), 1–11 (2012)
15. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
16. You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 29 (2015)