
Master Thesis



Tom Eijkelenkamp

Supervisors:

Dr. Almila Akdag Salah

Dr. Jiapan Guo

Abstract

In this thesis we study how to improve knowledge and skill of AI art generators on composition aesthetics. Positioning elements within a design is used in art to create visual effects, but this can be hard to achieve using diffusion models. We look at various methods to automatically analyse topics of composition in artworks, in order to use these as training data for diffusion models. We perform a more detailed study on three vision language models Llava, Cogvilm and Deepseek-vl. Their applicability for the task was evaluated in a survey involving students from art studies. Further analysis was done by generating training data to retrain Stable Diffusion and compare whether composition effects are easier to create. VLM's are versatile because of their natural language interface and therefore can be prompted to do many image analytics tasks. We find that VLM's rank differently in analysing per topic of composition and therefor it pays off to find the highest ranking model for each particular task.

1 Introduction

Composition in art and graphic design involves many topics on which an artist has to gain his or her skill to be recognized by the audience [34][30][17]. You have a certain goal behind your work and because of this you need to make use of certain effects to achieve the desired outcome. Composition is about how you make the layout of your design; what elements do you use and how do you arrange these throughout your design. Many studies have been done on what makes a composition successful, exploring what kind of effects can be used to ensure that the composition effectively communicates its intended message or evokes the desired emotional response.

AI systems that are made to generate visual art are modelled by millions of example images. They can render works that fall within the whole distribution that underlays the training images. The set which is used to train the AI with can contain all sorts of examples ranging from photographs, graphic design, fine-art works, anything that can be captured in images. It might contain work of a professional artist that had put thought into the composition, or just a photograph from a regular person that happened to capture a moment in time. AI art generated do not always understand or listen well when prompting them to render visual effects using composition. In the Figures 1 to 4 examples are shown prompting Stable Diffusion to generate images that visualize such effects, but the resulting images fail to capture the intended arrangement.



Figure 1: Result of prompting Stable Diffusion with: "An apple to the right of an pear".



Figure 2: Result of prompting Stable Diffusion with: "A pile in which apples stand out from pears by their contrasting colors".



Figure 3: Result of prompting Stable Diffusion with: "Apples that are laid out in a manner that draws the eye of the viewer from the foreground to the background".



Figure 4: Result of prompting Stable Diffusion with: "One pear that immediately draws the attention of the viewer among many apples".

In this thesis we study how to improve the knowledge of AI art generators on composition theory. We look at how to train diffusion models to have better compositional skills.

The questions that we investigate are: what part of the diffusion model architecture is best to look at for improving its ability to generate artistic composition; what algorithms to use for automatically generate training data for diffusion models that encompasses concepts of art theory; how well do these algorithms perform doing the task.

Our contributions to the field is that we infuse art theory of composition in the field of image generation.

First we discuss technical background of AI art and parts of the architecture that involve composition as well as art theories around composition. In section 3 we explain our method to our research, give technical details of possible algorithms that can be used, as well as a qualitative evaluation on these algorithms. In section 4 the set up for a human evaluation on a set of approaches is explained. In section 5 we show how we experimented with retraining an open source diffusion model with generated training data. Section 6 is an overview on the results of the experiments. This is followed up by a discussion section on our study and a conclusion.

2 Scientific background

2.1 AI Art

Art can be created using many methods and tools. Over time there has been many developments in the spectrum of styles and ideologies. Examples of art created by materials in the real world are drawings, paintings, sculptures and architecture. At some point methods became available to create art digitally, making it possible to visualize graphics on light emission screens. The computer can be used to efficiently compute geometric forms as well as many other mathematical patterns. This creates a whole range of new opportunities. Applications as three dimensional sketching, animation and motion design are much simpler using computers. All kinds of automation can simulate artistic patterns in large scale. For example self generating landscapes used in game design. The whole game industry is full of digital art. Now that AI has come to evolve more and more, people started to use machine learning techniques to synthesize the creative mind.

2.2 Diffusion models

One of the techniques that is used to automatically generate visual art is modeling by diffusion. A paper from Ho et al. [14] shows one of the earliest studies in this area. In order to be able to generate an image, the system is shown how to do so in examples. The creation is a process over many steps going from a noise vector (image) to a realistic looking picture. One example in the training process is an image that is diffused with noise over a certain time schedule, illustrated in Figure 5.

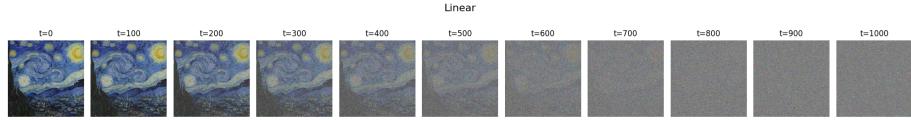


Figure 5: Example of an image interpolating between original content and complete noise. Here a linear schedule is used.

The model learns the reverse of the image diffusion. By examples is learned to how to predict and remove noise, and thereby transform the noise vector into a real looking image. The model is shown examples over the whole range of time steps.

This image synthesis process is very different from what a real artist is doing. In most cases an artist starts with a blank paper on which he or she sketches lines, adds colors or elements. Usually this is a sequential process of one modification at the time. The diffusion model learns to modify the whole images at once, removing noise from each segment in the image, and so creating all the elements in the picture in parallel, over a series of steps. The noise removal procedure can be related to the workflow of an sculpting artist. Here you start out with a stone in a random form, similar to a noise vector. Step wise you remove the artist chips the randomness away from the stone, and over time it starts to look more and more as a figure from the artist imagination.

The mathematics behind the training procedure of the diffusion module follows a regular supervised flow. In the training routine, the diffusion model learns to predict the noise at a time step. This prediction is compared to the true value (target) from the example. The magnitude of the error is used to tune each parameter of the model slightly in the appropriate direction to ensure the models prediction better matches the example value.

Many examples are shown to the model in the training phase to make the model learn the creation process of a whole distribution of visuals.

2.3 Noise schedule

We want to create a trajectory from a completely random pixelated image to an original image from the training data. Along this trajectory this random image transforms into the original image according to a certain schedule. This is simply defined by interpolating between the random image and the original. An image at a certain point along the trajectory is a ratio between random and original image. The schedule defines how this ratio changes along the trajectory.

A random image is defined as follows. The image has the same dimensions as the training image for which we want to create the trajectory. Each pixel in this image has a value with a certain variance around a mean. The value is Gaussian distributed and independent from any other value in the random image.

Different schedules can be used to change the signal to noise ratio along the

trajectory. You can use for example a linear or cosine schedule, illustrated in Figures 5 and 6.

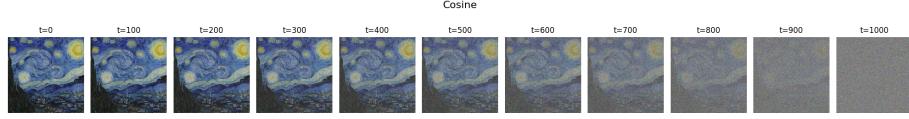


Figure 6: Example noise schedule, according to a cosine schedule.

For each time step the more noised image x_t can be calculated from the last step x_{t-1} using the forward process formula q :

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

Here β_t is the variance of the gaussian noise at time step t . Ho et al. use a linear schedule ranging $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.

The formula can be reformulated in order to directly go from time step 0 to t :

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \quad (2)$$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad (3)$$

$$\alpha = \sqrt{1 - \beta_t} \quad (4)$$

In the training loop a batch of images is taken from the data. These get noised to step t . The network predicts what noise used, after which the error on this prediction is used to compute the networks gradient in respect to the parameters. This is then used to tune the parameters accordingly.

Algorithm 1 Training [14]

```

1: repeat
2:    $x_0 \sim q(x_0)$ 
3:    $t \sim \text{Uniform}\{1, \dots, T\}$ 
4:    $\epsilon \sim \mathcal{N}(0, I)$ 
5:   Take gradient descent step on  $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}, \epsilon_t)\|^2$ 
6: until converged

```

2.4 Prompt

When the above described model is used to synthesize new images the user has no control over what kind of images are created. The model is only conditioned

to output an image that looks similar to the training images. There are several methods to guide this ‘sculpting’ process to create specific visual content. They allow the user to steer the image generation with a text prompt with various degrees of freedom. Some methods allow the user to ask for a image from a particular class, others give the option to give a detailed text in natural language and so a complete desired composition can be described in detail.

Classifier guidance: An initial approach that was applied is classifier guidance [8]. This uses a classifier that was trained separately on noised images. While the diffusion model is creating an image, this classifier predicts the class of that image. The information is used to steer the creation into the direction of a desired class.

Classifier-free guidance: Ho et al. [15] achieves a similar effect without the need for a separate trained classifier.

Clip-score guidance: In a study by Kim et al. [18] the text conditioning is implemented by adding a Clip-score [32] into the loss function. In CLIP, a text encoder and an image encoder are pretrained to identify which texts are matched with which images in the dataset.

Cross attention: Stable diffusion [35] uses attention [44] to implement text guidance. At the training phase the system learns a relation between visual aspects in the training images and words in natural language. Each training image example comes with a description that states what is seen in the image. Through cross attention the model learns what visual tokens in the example image relate to what text tokens from the description.

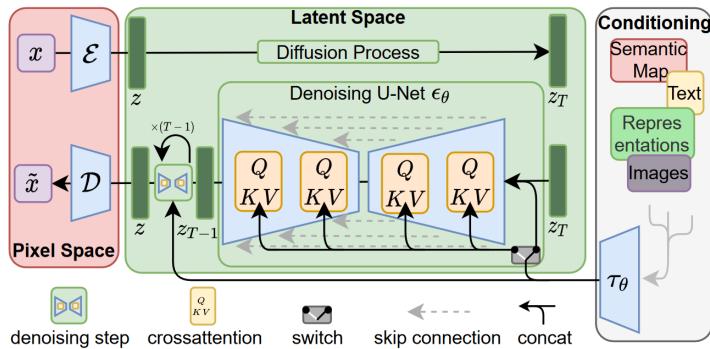


Figure 7: Architecture of latent diffusion model using cross-attention on the intermediate layers of the u-net. Adapted from [35].

Multiple modalities can be used to condition the image generation, one of which is language prompts. These are pre-processed with a domain-specific

encoder τ_θ that projects y to an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$. Stable diffusion as many other diffusion models use UNet [36] as neural network architecture. The intermediate representation of the conditioning is mapped to the intermediate layers of the UNet via a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$, with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y).$$

Here, $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon}$ denotes a (flattened) intermediate representation of the UNet implementing ϵ_θ , and $W_V^{(i)}$. $W_Q^{(i)} \in \mathbb{R}^{d \times d_\epsilon}$ & $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ are learnable projection matrices. See Figure 7 for a visual depiction.

Based on image-conditioning pairs, the conditional LDM is learned via

$$L_{\text{LDM}} := \mathbb{E}_{(x,y), \epsilon \sim \mathcal{N}(0,1), t} [\|c - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2], \quad (3)$$

where both τ_θ and ϵ_θ are jointly optimized via Eq. 3.

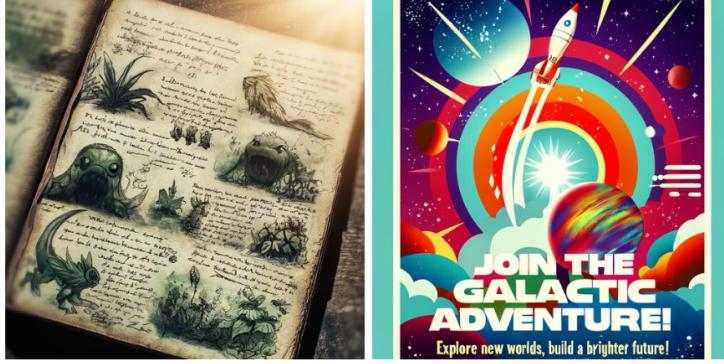
2.5 Training data

The studies DALL·E 3 [3] and Yeh et al. [47] show that improving the quality and detail of the training data captions, makes a diffusion model learn a rich and complex relation between visuals and texts, resulting in a model that can produce compositions that are closely aligned to the prompt. Big datasets such as LAION-5B [37] contain image captions that are rawly fetched out of the HTML in the web where the images were found. This often lacks detail or contains inaccuracies. Figure 8 illustrates an example used in the paper from OpenAI of different levels of detail in image captions. They use an algorithm to create detailed captions for their DM training data. Their model trained with such rich detailed captions listens much closer to what kind of composition is asked for in the prompt. When training the system with comprehensive descriptive text for each image, it requires a prompt of similar detail when generating images. To solve this issue they use their language model to upsample short prompts to a highly detailed ones. Examples of generated images on several complex prompts from their study are show in Figure 9.

	<p>Image</p> <hr/> <p>Alt Text</p> <p>is he finished...just about!</p> <p>23 (19 of 30) 1200</p> <hr/> <p>SSC</p> <p>a quilt with an iron on it.</p> <p>a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.</p> <hr/> <p>DSC</p> <p>rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.</p>
---	---

Figure 8: Examples of three levels of detail in captions, alt-text from the HTML, short synthesized caption and descriptive synthesized caption. Adapted from [3].

As a evaluation metric they used Clip score [13] on generated images and the related prompts. This creates an embedding for the image and prompt using the Clip model, and calculates a cosine similarity between the two embeddings. The score has a minimum value of 0 and ranges from 0 to 100, where 100 corresponds to complete alignment between prompt and image. They outperform stable diffusion xl (30.5) and dalle 2 (31.4) on Clip score with a score of 32.0.



Ancient pages filled with sketches and writings of fantasy beasts, monsters, and plants sprawl across an old, weathered journal. While the faded dark-green ink tells tales of magical adventures, while the high-resolution drawings detail each creature's intricate characteristics. Sunlight peeks through a nearby window, illuminating the pages and revealing their timeworn charm.

A vibrant 1960s-style poster depicting interplanetary migration, with a retro rocket ship blasting off from earth towards a distant, colorful planet. Bold typography announces "Join the galactic adventure!" with smaller text underneath reading "Explore new worlds, build a brighter future." The background features a swirling galaxy of stars and constellations.

Figure 9: Examples of generated images after training a diffusion model with training data where the image captions are synthesized in high detail. These results come from the development of DALL·E 3 by OpenAI. Adapted from [3]

2.6 Starting noise

The study of Mao et al. [28] finds that content that is being generated at a certain location in the generated image is dependent on the starting noise at that location. The attention maps at those locations at the start of generation find noise at this location to relate to patterns of particular objects. It is found that therefore some content are very easy to generate starting from this initial noise, but other content is very unlikely to generate. The attention maps are very important for what the arrangement is in the resulting image. The values of the initial noise is important to what content is generated, not the location. Therefore this noise can be moved around to modify the image composition. In their study they come with a method to modify the initial noise in order to get a certain arrangement in the image, visualized in Figure 10.

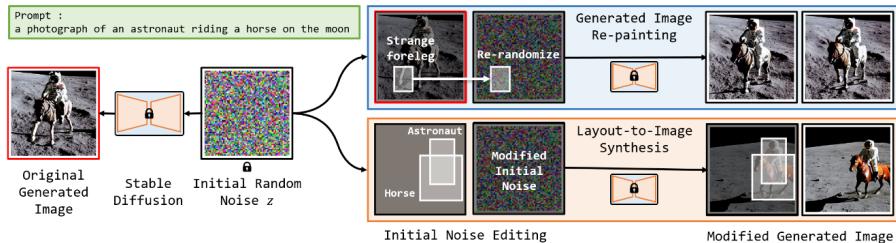


Figure 10: In this study, they investigate the impact of the initial image on image generation and propose a novel direction for controlling the generation process by manipulating the initial random noise. Adapted from [28].

2.7 Image editing

Region modification: Other than guiding the trajectory with a text prompt to create an image with a certain composition, the interface can be extended by adding a region modification tool. A generated image can be altered subsequently by selecting a specific region and which is then modified in line with a text prompt [12][29][2]. An example from the study of Avrahami et al. can be seen in Figure 11.

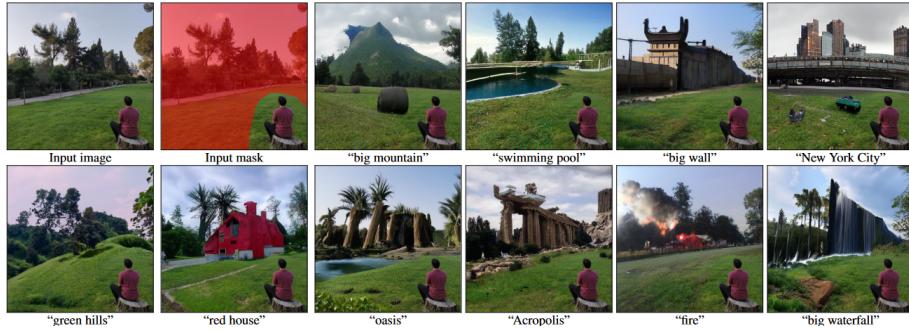


Figure 11: Modifying a specific selected region with a text prompt. Adapted from [2].

Certain modifications are not possible using this method, such as changing the texture of an object, since the targeted region is removed from the image in the process.

Prompt modification: The study by Hertz et al. [12] uses the cross-attention layers in order to do small adjustments to a generated image. They allow the prompt to be altered after it was used to generate an image. Only the adjustments to the prompt are used to alter the image, the attention maps related to the rest of the word tokens are frozen. The Figures 12 and 13 show what this looks like in practice.

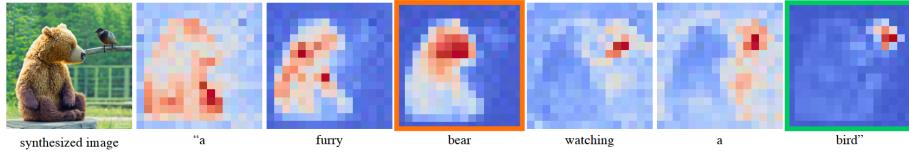


Figure 12: These figures illustrate activation levels at regions in the attention map for given words from the description. An average across all the diffusion time steps is shown. You can see that every word relates to a region in the image that visualized that meaning. Adapted from [12].

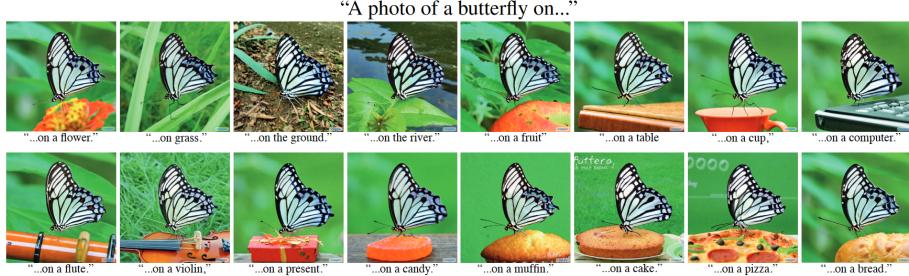


Figure 13: Object preservation. By injecting only the attention weights of the word “butterfly”, taken from the top-left image, we can preserve the structure and appearance of a single item while replacing its context. Note how the butterfly sits on top of all objects in a very plausible manner. Adapted from [12].

2.8 Latent space

The article [35] finds that by using an encoder and decoder to perform the computations on a latent version of the training images, reduces much of the required computations while retaining the realistic quality of images. The encoder creates a compressed version of the training images, which are then noised in this latent format. The decoder transforms a generated latent image into image format again. The resulting architecture change is visualized in Figure 7. This enhancement improves the diffusion model in its capability to generate a composition, since more easily a model can be trained with large amount of data.

2.9 Self attention

Several works of diffusion models also use self-attention [44] on the visual domain [14][8], implementing it within the UNet [36] structure. This relates any visual pattern to another within an intermediate feature map at a certain step in the trajectory. This attention mechanism has a much wider range than the small local view of a convolution kernel. This therefore goes without the assumption of local feature mapping. Specifically, for the height H and width W , given any feature map $X_t \in \mathbb{R}^{(HW) \times C}$ at a timestep t , the N -head self-attention is defined as:

$$Q_t^h = X_t W_Q^h, \quad K_t^h = X_t W_K^h, \quad (5)$$

$$A_t^h = \text{softmax} \left(\frac{Q_t^h (K_t^h)^\top}{\sqrt{d}} \right), \quad (6)$$

where $W_Q^h, W_K^h \in \mathbb{R}^{C \times d}$ for $h = 0, 1, \dots, N - 1$. Each A_t^h is then right multiplied by $V_t^h = X_t W_V^h$, where $W_V^h \in \mathbb{R}^{C \times d}$. Moreover, diffusion models using Transformers [44] as the backbone has also been proposed [31].

2.10 Relation to the process of real life painting

An artist in real life creates a piece by which he or she also undergoes various stages. Usually first a global lay out is sketched that illustrates the foundation of the composition. This phase can be related to composing the initial noise in Mao et al. [28]. For a three dimensional scene, you continue on this by first painting elements that are farthest away, where after you add step by step elements that are each time a little closer to the viewer. The most distant pieces are depicted in less detail and this detail smoothly is increased for pieces closer by. You usually do not change the overall composition, but mostly fine tune what you laid out.

2.11 Composition

When you create a visual, you can design a certain composition that it must follow. What this means can be studied from many angles. Artists have different perspectives and styles on the subject. In general it is about how you layout all the parts of your design. How do you position all the elements in your work. People have theories on how to create certain effects and have perspectives on what looks good.

Looking at it from a basic geometric view every object, form or shape has a particular location in a picture. This can be shown in either two or three dimensions. It has a certain size and it is rotated towards a direction. In addition to this the viewpoint also is located likewise at a location viewing the scene from a direction. The view port could be orthogonal or perspective wise. Figure 14 illustrates this in two examples.

On a more abstract level you can make layout follow a certain pattern. Edgar Payne talks about this in his book on composition of outdoor paintings [30]. The overall layout can for example form a triangular shape, a circle or a curve. Some examples are shown in Figure 15. These examples are telling something about how all of the elements as a whole are forming a particular shape. These are all pictures of landscapes, but it can be applied to any kind of design.

Using rhythm you can create effects and styles for your composition. By repeating shapes or using a particular movement in brush or pencil strokes you can create many different effects. In Figures 16 to 20 show examples of patterns you can create by playing with rhythm.

¹<https://nl.pinterest.com/pin/376261743851180155/>

²<https://nl.pinterest.com/pin/46133755590353370/>



(a)

(b)

Figure 14: In these two posters (a)¹ and (b)², you can see that each element has its specific location, size and rotation. Partially there is a layout in two dimensions for the headlines and descriptions. Each text has a clear location and size. The three dimensional pictures show a more complicated scene of elements. Here you have the additional depth dimension and so parts can be behind or in front of other. Mostly all the elements are facing a regular direction, the guitar on the back of the person has a slightly tilted rotation.



Figure 16: Regular rhythm³: predictable repetition of one or more elements.



Figure 17: Random rhythm³: irregular placements of one or more elements.



Figure 18: Flowing rhythm³: Continuous movement of one or more elements in a smooth organic manner.



Figure 19: Alternating rhythm³: two or more elements are repeated in a predictable sequence.



Figure 20: Progressive rhythm³: Gradual changes of one or multiple characteristics of the repeating elements.

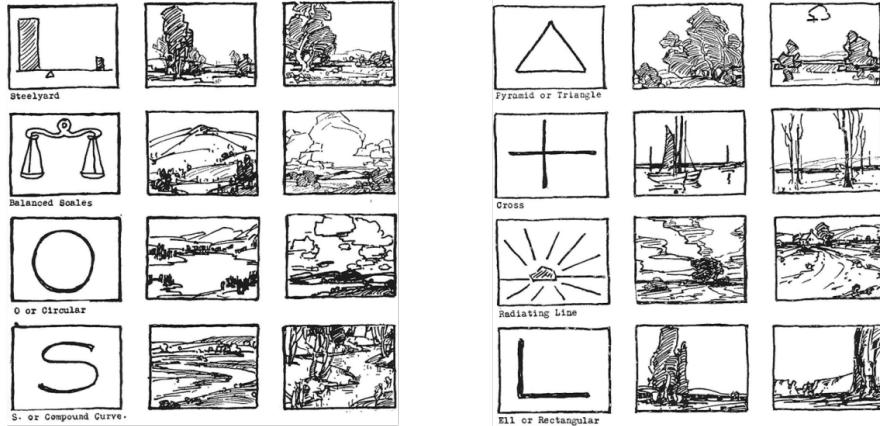


Figure 15: These figures illustrate how the overall picture's arrangement is following a particular shape. Each shape has particular characteristics, such as creating a balanced layout. Adopted from Payne's book [30].

Furthermore you can describe how all of the elements are laid out across the design by the terms unity, balance and contrast. To describe the balance in an image the book [34] give very nice examples, here shown in Figures 21 to 23. The author of the book relates balance within pictures to how humans feel a sense of unease when there is a lack of balance. We have symmetrical bodies and when we lose balance we fall. Balanced images make you feel calm and settled, while unbalanced ones suggest precariousness.

These terms seem more subjective, the motivation why elements are contrasting each other or form balance can be the opinion of the artist. It might not be immediately clear to anyone why a certain arrangement forms a unified whole. The artist might have an idea behind why he made a certain set up, why certain elements are contrasting or forming a balance. This is a thought formulated in language or it might also only show as a visual. There is balance in the form, since there are only round shapes. There is contrast two persons show in a picture, since they are leaders from different world countries.

³<https://artclasscurator.com/rhythm-in-art-examples/>



Figure 21: Greg Hunt and Dylan Rieder, photographer Anthony Acosta, 2008. Expressed as an equation, this image would be: $2 \times \text{empty space} + 1 \times \text{active space} = \text{perfectly balanced image}$; it devides up into horizontal thirds. The narrowest strip at the bottom of the photo is also visually busiest; it's where all the action is happening. The large expanse of green wall serves as respite to the flowing motion and activity below; it's a place where a viewer's eye has an opportunity to take a break. Adapted from [34].



Figure 22: Private Soldier, Battle of Waterloo, 1815, photographer Thom Atkinson, n.d. In his series Inventory, Atkinson laid out and cataloged all the possessions of single soldiers from various time periods in history. An image like this could easily feel jumbled or messy, but the meticulous attention paid to balancing groups of small objects with larger ones provides a soothing sense of order and balance, allowing a viewer to appreciate each item at leisure. Adapted from [34].



Figure 23: Photographer Mike Krautter, n.d. Sometimes visual balance is achieved by placing one element little off-kilter. The glass at the far right, whose rim juts over the edge of the concrete and interrupts its straight trajectory, provides a focal point midway between the glasses, their beautiful shadows, and the dancing blue water of the swimming pool. It also feels a bit precarious, peeking over the edge. Adapted from [34].



Figure 24: In the left picture³ all the colors all fall in the same scheme and the characters seem to belong together enjoying the same activity. In one of the cover arts of the band Joy Division⁴ shown in the right picture the text is cut up and misaligned. These two pictures show a composition of unity and it's opposite.

3 Methodologies

In order to implement a diffusion model with knowledge about art theory we look at several algorithms to generate training data that reflects art theory. Training data for diffusion models consists of image text combinations, where the text reflects what can be seen in the image. We use vision models on art works and evaluate how well they can be used to create the required textual data. We do a qualitative evaluation in order to choose a set of vision models which we are going to study in more depth. Human evaluation is set up to perform an in depth analysis of this set of vision models. On top of this we generate training data using the vision models and retrain an existing open source diffusion model with such datasets. The resulting new diffusion models we test on how well the generate artworks following prompts that contain tasks about art theory concepts. We do a quantitative evaluation on generated images by measuring the Clipscore on the prompt image combinations.

³<https://nl.pinterest.com/pin/508836457868951688/>

⁴<https://nl.pinterest.com/pin/35184440824947272/>

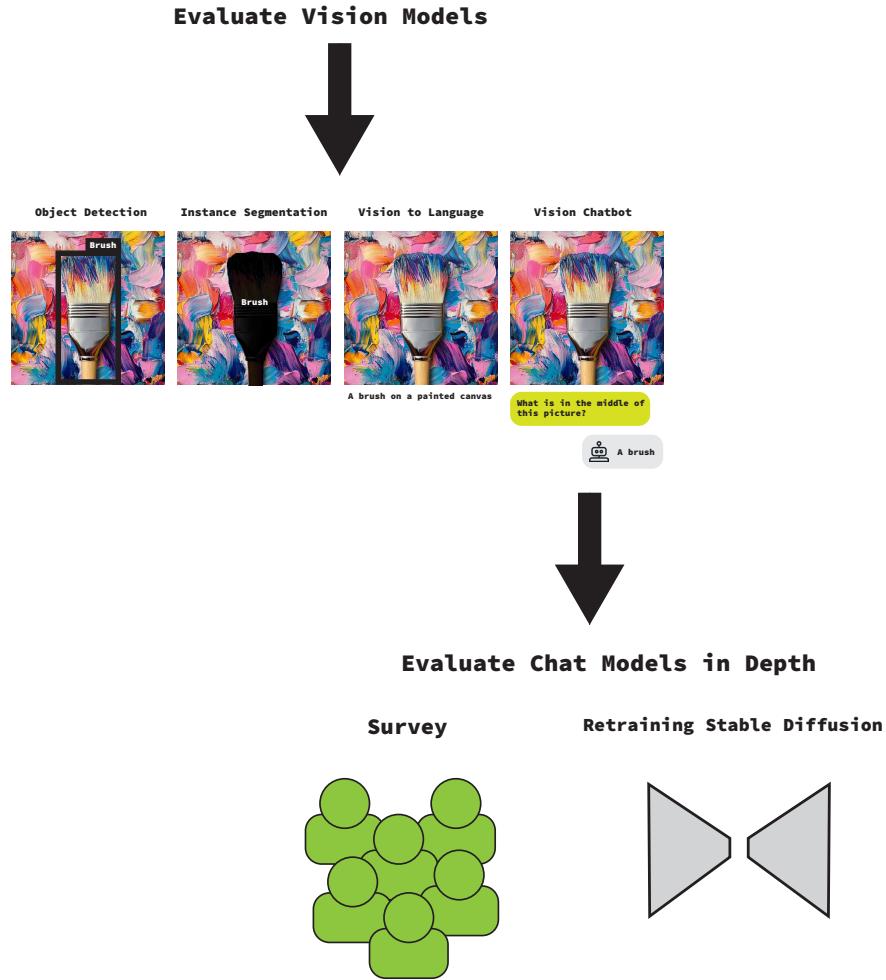


Figure 25: Thesis research pipeline. We started with a literature review on what part is best to focus to improve composition skills of diffusion models. Continuing on evaluating methods to infuse art theories into training data, looking at various vision models. The main focus of this research is an in depth analysis of Vision language chat models, by means of a survey as well as experiments on retraining an open source diffusion model.

3.1 Vision models

The following methods could be used to automatically describe various compositional concepts.

Object detection or image segmentation can be used to localize objects and the area they provoke. In this sense the most basic ideas of composition could

be described, including proportions. Even balance in layout could be computed by a mathematical formula using the detected locations of objects. Whether an object is the focus point of the composition or a topic such as contrast could be more difficult to compute using these particular vision methods.

Yolo: This is a famous one-shot object detection network [33]. It can predict bounding boxes and corresponding classes for multiple objects in an image in one pass through the network. This is done by dividing the image into partitions and predicting the probability of the presence of an object and the bounding box coordinates of the object, as illustrated in Figure 26. It also predicts the class of the object. The architecture is a convolutional neural network. The architecture is improved over many versions of the network, making it more accurate, faster and better at detecting objects at different scales.

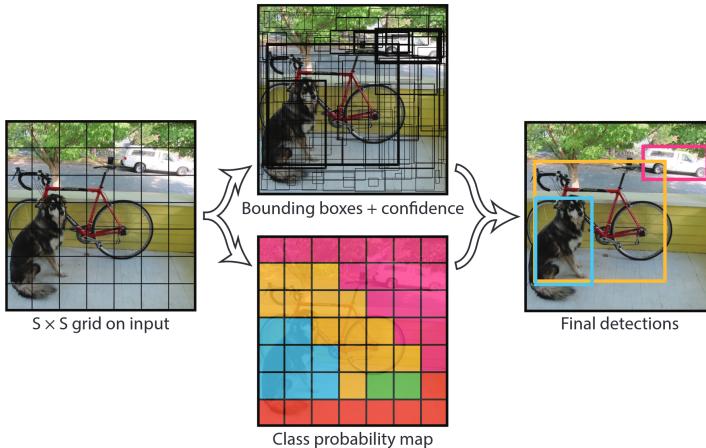


Figure 26: Images are partitioned into a grid in the Yolo network. On each segment a bounding boxes plus a confidence level on the presence of an object in this region. Additionally class labels are predicted for each segment. This all is done in one pass through the network.

Yolo version 8⁵ can be downloaded trained on the Coco dataset [23] including 80 possible to detect classes, as well as trained on ImageNet [7] where it can detect 1000 different classes.

The following images in Figure 27 contain results of the Yolo version 8 network performing object detection on a set of bandposters⁶. This version of Yolo can perform a number of tasks: image classification, object detection, instance segmentation, oriented bounding boxes and detection of body poses.

The detection and segmentation does not work very well on the band poster

⁵<https://github.com/ultralytics/ultralytics>

⁶<https://www.collectionzz.com/category/concert-posters/database>

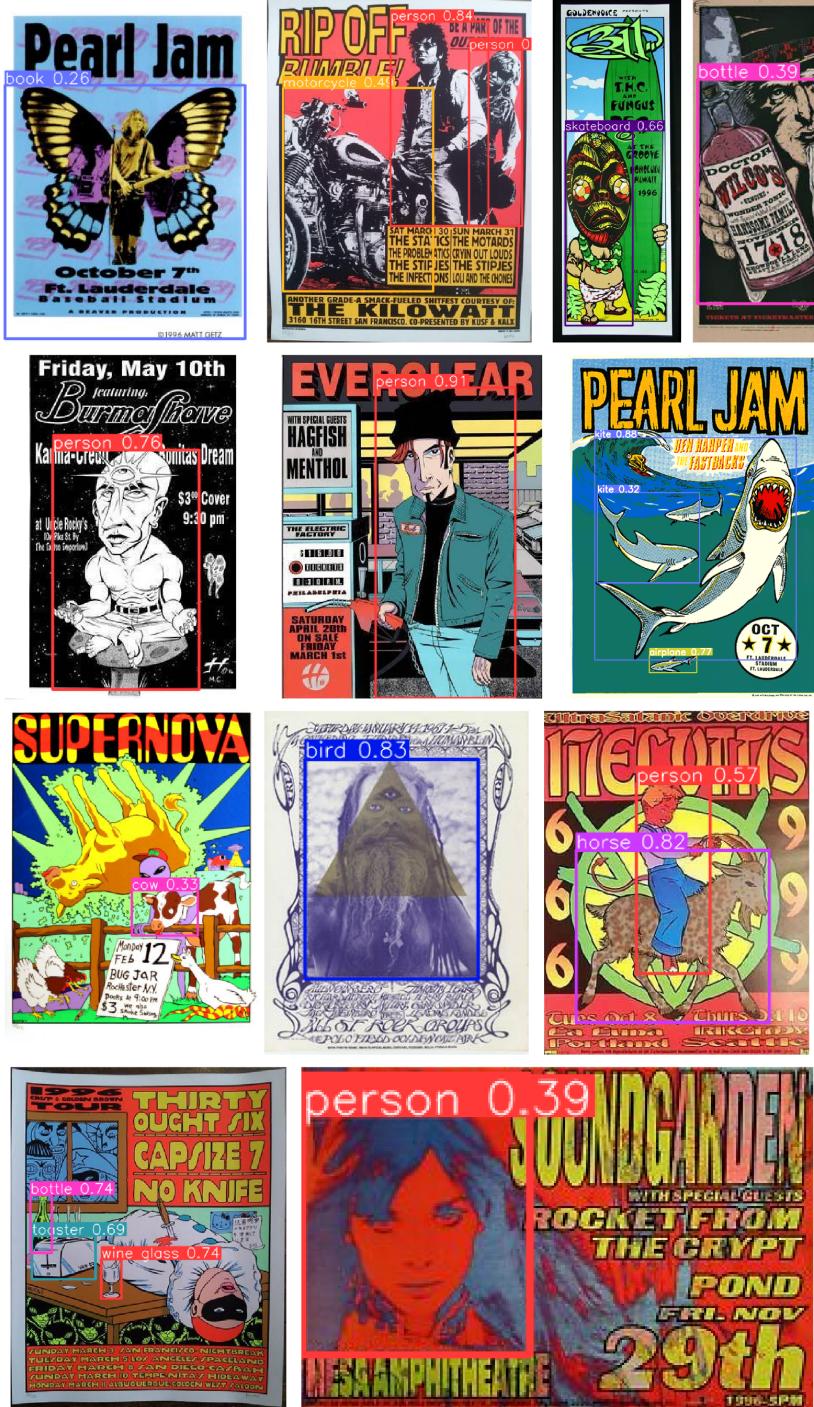


Figure 27: Results after applying Yolov8 object detection on a set of band-posters⁶.

set. The test can be improved by experimenting with other input dimensions, although results might not be much better than running with default setting. Poses work well for detecting persons, which seems to work better than person detection. Oriented bounding boxes give very bad results.

SAM: This network is developed by Meta and is used for segmenting object in images[19]. Input can come in a variety of ways: you can give a point in an image and the object at that point will be segmented; you can input a prompt and the image will segment accordingly; or you can segment the image into all its parts.

The model consists of three parts: a powerful image encoder computes an image embedding, a prompt encoder embeds prompts, and then the two information sources are combined in a lightweight mask decoder that predicts segmentation masks.

Points and boxes are represented by positional encodings [42] plus learned embeddings for each prompt type (point, box, mask) and free-form text using text encoder from CLIP [32].

Images are encoded with a slightly adapted version of MAE [10] pre-trained Vision Transformer (ViT) [9], to process high resolution images.

The mask decoder is a modification of a Transformer decoder block [44] followed by a dynamic mask prediction head. It uses self-attention on the prompts and cross-attention in both directions, prompt-to-image and image-to-prompt.

A prompt could be ambiguous in the sense that a point on a shirt could refer to that shirt or the person wearing it. To deal with this the model is trained on multiple reasonable mask options per prompt. It will predict various masks and their confidence scores.

An overview of the model is depicted in Figure 28.

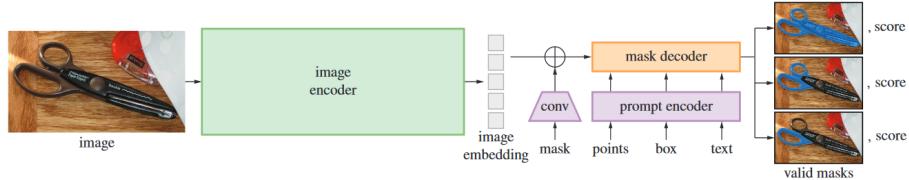


Figure 28: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores. Adapted from [19].

AutomaticAnnotation: This is an algorithm that uses both object detection by localizing it within a bounding box as well as segmenting the object with a segmentation mask [45]. The framework is built on top of an implementation [1] of Mask R-CNN [11]. It was trained on the Coco dataset [23] and therefore

it by default predicts such a set of object classes. We tested this tool on a set of bandposters⁶ from which some examples are shown in Figure 29.

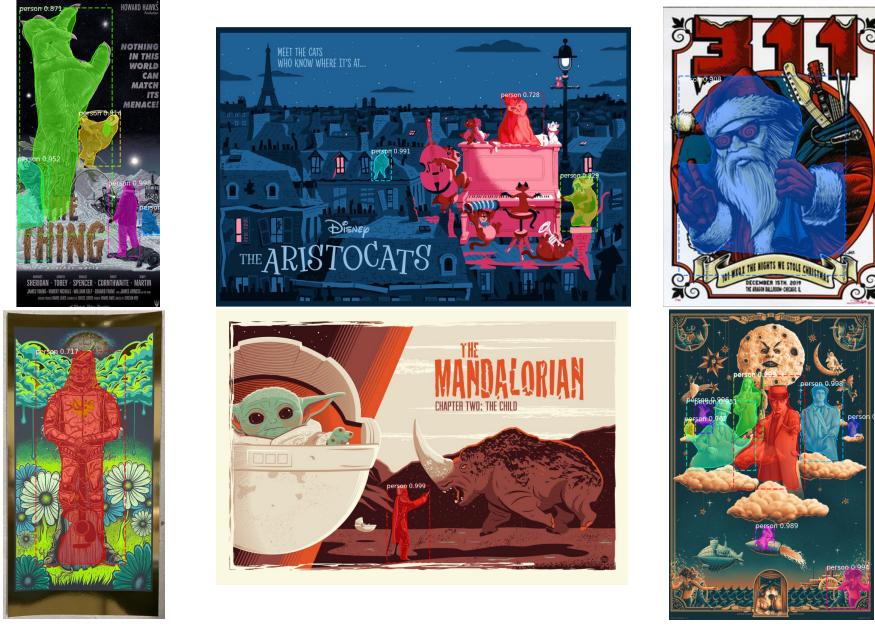


Figure 29: Examples of usages of automatic annotate on a set of bandposters⁶.

3.2 Vision Language models

Since large language models show there broad range of capabilities, people have been working on developing this same functionality in the multi modal field. In a similar fashion they trained models with instructions that involved both visual and text features. This way you can set up a chatbot that you can input an image along with an instruction and have the AI solve the task. You could for example ask the AI to give a description about this image. In this manner you could set up a pipeline to create image-text training data.

Blip captions: This computes a short text description for an image [21]. As input you give it an image and it will compute the description. There is no way of controlling what the description is about. This makes it less attractive for using it to describe composition, since it will most likely not describe subjects as focus point or balance and contrast. Examples of what Blip captions look like on artworks from the WikiArt dataset [8] are shown in Figure 30.

VisualBert: A vision language model that can be given an image along with a natural language prompt[22]. The model outputs from a list of preset one word answers.

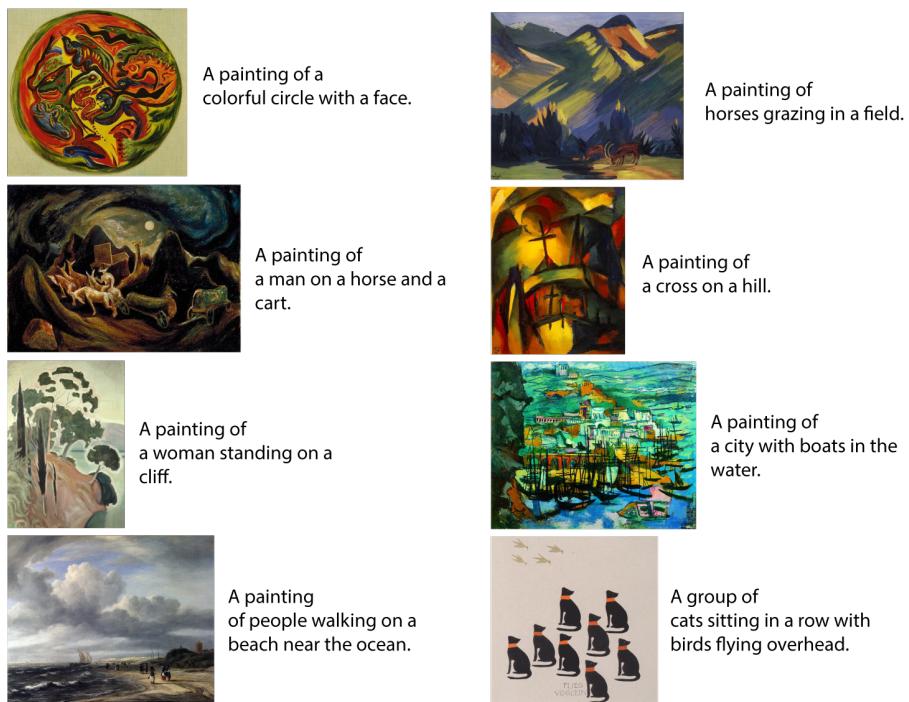


Figure 30: Blip captions for paintings selected out of the WikiArt dataset [8].

MiniGPT: This is a vision language model that also can be given an image plus a prompt in natural language[49]. The model responds in a short paragraph on this input.

Chatgpt/gpt-4o: A proprietary language model that comes with vision capabilities. You can input this model with an image on which you can have a conversation. The responses are of good quality in terms of detail and accuracy. It is not open source and therefore money need to be provided to access the model and you are reliant on the computational resources of the provider.

Llava: This is a vision language model that has a chat interface[25][24]. They use a pretrained large language models Vicuna v1.5 13b [5] and enhance this with a CLIP visual encoder ViT-L/336px [32] which is used to create an embedding for images, that is then projected to have an appropriate number of features with a multilayered perceptron. The architecture is visualized in Figure 31.

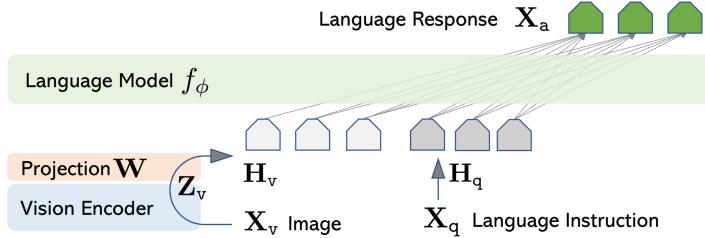


Figure 31: The architecture of Llava. It is trained with image instruction answer combinations. Images get encoded through a vision encoder and a projection layer is used to scale the number of features to align with the number of features of the instruction. Adapted from [25].

The system is trained on generated instruction data. First 595k image-text pairs are used from the CC3M dataset [4]. For an image and its associated caption, it is natural to create a set of questions with the intent to instruct the assistant to describe the image content. They prompted GPT-4 to curate such a list of questions, which are used at the first stage of training.

Thereafter the system is finetuned on 158K examples from the Coco dataset [23]. These are used to create three types of conversations: conversation, detailed description and complex reasoning, examples of which are show in Figure 32.

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

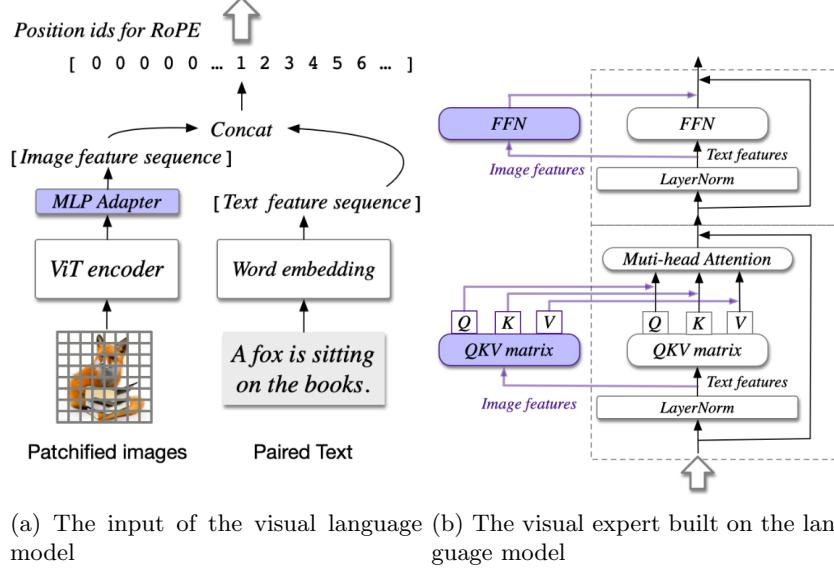
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Figure 32: Examples of generated instruction conversations for Llava[25] training data, using the Coco dataset of caption and boundingbox information and transforming this information using gpt-4.

The evaluate their system on the ScienceQA benchmark [27], the first large-scale multimodal science question dataset that annotates the answers with detailed lectures and explanations.

Cogvlm: Similar to Llava, this is also a vision language model enhanced with a chat interface [46]. Here they use the pretrained Vicuna v1.5 7B large language model [5] which they enhance with a visual expert module doubling the parameters in the system. For image encoding they use EVA2-CLIP-E [40] and scale the feature dimensions using MLP adapter [38]. For each layer in the large language model an visual expert layer is added consisting of a QKV matrix and MLP, identical of those in the language model, illustrated in Figure 33.



(a) The input of the visual language model (b) The visual expert built on the language model

Figure 33: The architecture of CogVLM. (a) The illustration about the input, where an image is processed by a pretrained ViT and mapped into the same space as the text features. (b) The Transformer block in the language model. The image features have a different QKV matrix and FFN. Only the purple parts are trainable. Adapted from [46].

They claim that this allows for a deep fusion between image features and features captured in the language model attention heads. This differs from shallow alignment methods represented by InstructBLIP [20] and MiniGPT-4 [49], where they connect a frozen pretrained vision encoder and language model via a trainable Q-Former or a linear layer, mapping the image features into the input embedding space of the language model. They say this deep connection using the visual expert layers allows for more complex reasoning. During training they keep the original language model parameters frozen, keeping the language capabilities intact. When this is not done the system is vulnerable for catastrophic forgetting. The setup is inspired by Lora training [16]. Due to their large architecture it requires a lot of VRAM on the GPU.

Deepseek: Their model [26] is built upon DeepSeek LLM [6] whose micro design largely follows the design of LLaMA [43]. They use SigLIP [48] for visual encoding in low resolution and SAM-B [19] for visual encoding in high detail. A MLP is used for visual feature adaptation.

In order to retain the language features during training they create an intricate training schedule that carefully balances the multi-modalities. They use an extensive list of training sets containing 4 sets of interleaved image-text data taking a ratio of 13.1% in the joint vision and language pretraining stage, 3 sets of image caption data (11.1%), 12 sets of table and chart data (2.1%), 2 sets

of web code (0.4%), 10 sets on scene text optical character recognition (1.2%), 1 set of document optical character recognition (2.1%), 1 set of text-only corpus (70.0%).

At the finetuning stage they use one set of in-house data (10.5%), 6 sets of general multi-modality data (35.5%), 3 sets of table and chart data (4.1%), 2 sets of web code (2.0%) and one set of text-only data (47.9%).

Qualitative evaluation: We tested the VLM’s Llava, Cogvlm and Deepseek on various datasets, for which we generated art analytics on numerous topics of composition.

Datasets. We selected paintings from the Wikiart dataset [41]. One set of paintings was a uniform distributed set over all the genres. Furthermore we selected two theme specific sets, one containing landscape artworks and another paintings from the renaissance period. On top of this we filtered a bandposter⁷ and graffiti⁸ dataset for a selection of evaluation examples. To add to that we used a selection of artworks from Pinterest for our qualitative evaluation.

Prompts. We asked the models to generate analytics on 22 different topics of composition, formulated in different ways, combining to 40 different prompts. The topics consisted of:

- General analysis of composition
- Unity
- Balance
- Make a list of all elements
- List all colors
- Movement
- Rhythm
- Focus point
- Contrast
- Patterns in the arrangement
- Proportions
- Foreground background relationship
- Figure ground relationship
- Open or closed composition
- Symmetry
- Rule of thirds
- Golden ratio
- Leading lines
- Gestalts principles
- Perspective
- Use of lines
- Eye movement

All the results can be seen in the web interface⁹. Examples of VLM responses

⁷<https://graffiti-database.com/>

⁸<https://www.collectionzz.com/category/concert-posters/database>

⁹https://photosynthesismembrane.github.io/caption_interface/

on a question to give a general description of the composition shown in Figures 34 to 37.

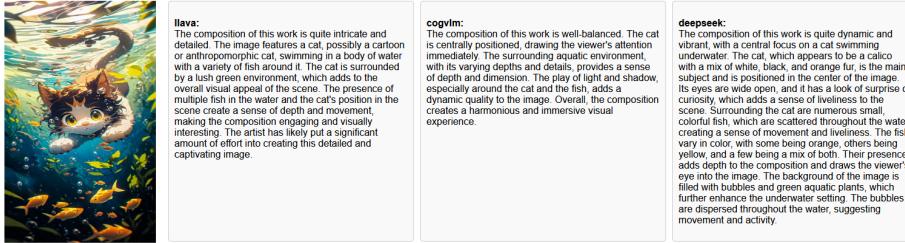


Figure 34: Example of VLM responses on artwork from Pinterest.

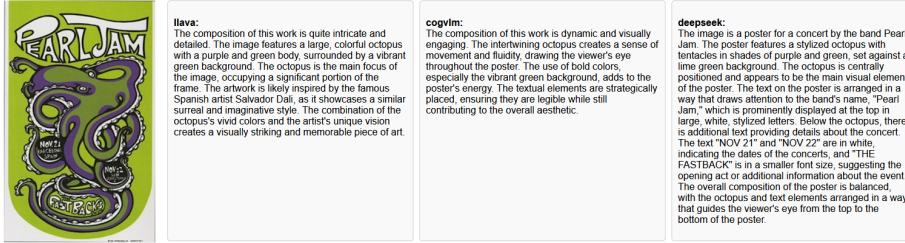


Figure 35: Example of VLM responses on a bandposter of Pearl Jam.

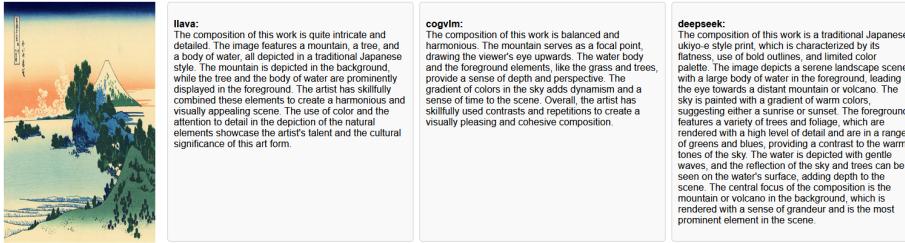


Figure 36: Example of VLM responses on artwork out of the WikiArt dataset.

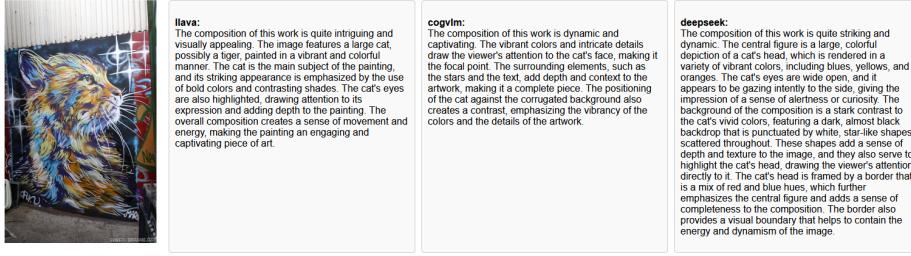


Figure 37: Example of VLM responses on a graffiti artwork.

Results. Overall you could see that the accuracy's of the answers depended very much on the type of artworks. Descriptions about landscapes were quite accurate as apposed to band posters and graffiti. These last two often contain unusual arrangements which are harder to describe. Especially the graffiti works, since they contain a lot of wildly stylized characters that the AI cannot interpret. It sometimes guesses what the letters are or what is written down, but this can be very wrong.

The responses can contain hallucinations. Descriptions of objects that are not there, although it would be a usual setting for these objects. For example the model would describe that there are ducks in the water, although the water is empty. The model might describe the sky as blue, although it is entirely red and orange by the sun.

Certain terms are not well understood. The model might use words as harmony and balance for a scene that depicts a setting of war or images that contains skulls and snakes.

Resource metrics. Comparing the time it takes to get results, Llava is almost two times faster than Cogvlm and Deepseek that run about equal speed. Cogvlm had to be set up with two a100 40gb gpu's due to it's size. The responses of Cogvlm were shorter in comparison, Llava responses were a bit longer. The answers from Deepseek were very extensive.

4 Survey

We want to evaluate how well the vision language models are suited for analyzing the composition of artworks. We do this in the form of a survey to get a human perspective on the AI's behavior. We specifically focused to get students from art studies to be a large proportion of the participants.

The goal of the survey is to find out what people find the best VLM to interpret composition of artworks. In more depth, we want to find if there is a significant difference between the VLM's on discussing the various topics of composition. Additionally we investigate what people find are the strengths and weaknesses of the VLM's for this use case. We are going to study the models Llava, Cogvlm and Deepseek.

We considered various approaches for the evaluation of these AI models. Three types of interfaces were considered from which two were used in the survey. We investigated several techniques on how to make the VLM's generate artwork interpretations. Furthermore we contemplated on what kind of statistical metrics to use.

4.1 Interface

We will show the participants artworks alongside with AI narratives on composition. The participants will give feedback on the quality of the narratives.

- This feedback could potentially given by showing participants an artwork along with one narrative, where the participant judges the quality on a likert scale. A downside of this approach is that the participant might get lazy and give an average score to all narratives.
- Another option is that we show an artwork and two narratives from two different VLM's, and have the participant judge which narrative aligns better to what is shown in the artwork. Here the participant need to pay close attention to what makes one narrative better over another. He or she must decide that either one of them is better and so it cannot give an average score on both narratives. He or she must find key features that indicate better quality. Statistical methods can be used to measure which VLM is favored by analyzing many of these side-by-side comparisons of the VLM's perspectives.
- As third option we could show the participant an artwork with one narrative and ask the participant to highlight parts of the text which do not fit or could be improved. This gives detailed information on what kind of mistakes the AI's make and what content is regarded as good enough.

4.2 VLM's narratives

Models: We generated AI narratives using the three different AI models llava-hf/llava-1.5-7b-hf, THUDM/cogvlm-chat-hf and deepseek-ai/deepseek-vl-7b-chat.

Prompt engineering: We want the VLM's to give a complete description on composition. As discussed in an earlier section such a description could comprise many subtopics, such as balance or use of contrast. We want a detailed view on the knowledge in these VLM's.

A simple approach is to ask the model to give a detailed analysis of the artwork's composition. This often lacks a description on many of the subtopics of composition.

Another way is to ask specifically to describe the artwork on each of these subtopics of composition. This creates a detailed, but extensive texts. It would be intimidating to a participant is asked to read such extensive texts.

We decided to prompt the model for descriptions only on one subtopic at the time. This keeps the texts reasonably short. For each artwork we show the participant a narrative on one subtopic, but we vary the subtopics over the different artworks. This gives a detailed analysis of the VLM’s performances.

Still the AI’s responses could be too long. We decided to truncate all the texts to consist of a maximum of 1000 characters.

In the prompt we considered describing in a short summary what the theory is behind the subtopic. We went to go with a short variant to simply ask the VLM’s to discuss the subtopic of the paintings composition. An professional artist advised to use keywords such as: ”The use of lines in the composition”, ”the use of foreground/background relationship in the composition”, ”the use of figure/ground relationship in the composition”, ”the use of symmetry/balance in the composition”.

The VLM’s painting analysis in the context of a subtopic often is explained using terms from other subtopics. This might have to do with that the chat version of the models was used, which might have in memory what earlier conversations were about.

Topics: We got advised by a professional artist to focus on particular subjects of composition. This included the list: ”Foreground-background relationship, figure-ground relationship, closed-open composition, gestalt principles. The use of lines, quality of lines, direction, pressure, width, etc. / How the eye moves along the composition generally helps us in a more precise description. Symmetrical balance or asymmetrical balance.”. Not all participants in the survey will understand these topics, because of this some of the topics are left out in order for everyone to be able to have a fair judgement. The final list of topics comprised: ”Composition”, ”Contrast”, ”Foreground & Background”, ”Symmetry & Asymmetry”, ”Eye Movement”, ”Focus Point”, ”Balance”, ”Proportion”.

4.3 Structure

Datasets: We made a selection of images to use for our human evaluation. For these we generated the AI composition analysis. The images are divided into two differently themed sets. One set consisted of renaissance paintings which were selected from the WikiArt dataset [41]. Two art experts advised to use paintings from the era renaissance, since this was an important time of development of composition in art. In addition to this set, we added another selection with a different theme, in order to study how sensitive the AI composition interpretations are to the what art genre the works depict. The other set is a selection of illustrations gathered from Pinterest. The two datasets can be found in the survey repository¹⁰.

Size: In order to make our findings of significance a number of results need to be gathered. This means that we need a minimum number of participants,

¹⁰<https://github.com/photosynthesismembrane/survey>

AI generated art interpretations and selection of artworks.

We are bounded by the number of participants we can get. We have a certain budget to give them which is 10 euros when the number of participants ranges from 20 to 50.

The survey ended up lasting 20 minutes. Earlier versions lasted longer but received comments for being too extensive, so it was shortened.

We selected 40 artworks per dataset. Per artwork we generated narratives for each of the 8 topics which adds up to 360 different narratives. That is per model, and since we used 3 VLM's in total that is 1020 narratives. The combination of the two datasets makes 2040 narratives.

The participant can chose the dataset for which he or she wants to do the survey. This gives some freedom to the attendant to do the task by his or her own preference.

The survey is divided into two parts as explained before, one part where one artwork is shown along with two narratives from different models, another part where one artwork is show along with one narrative.

(1/4) Contrast

Judge which description fits best

Judge which **text** you find that bests describes the use of **contrast** in the painting. Click to select the most fitting description.

In the painting, there is a contrast between the elements of the scene, which adds depth and interest to the composition. The main focus is on the group of people gathered around a stone table, with some of them sitting and others standing. The people are engaged in a conversation, which is the central activity in the painting. In contrast, there are two birds in the scene, one located near the top left corner and the other near the top right corner. These birds add a touch of nature and a sense of movement to the otherwise static composition. Additionally, there is a book placed on the table, which might suggest a shared interest or topic of discussion among the people. The combination of these elements creates a dynamic and engaging scene, with the people at the center of the action and the surrounding elements contributing to the overall atmosphere and narrative of the painting.



The painting depicts a scene with a mix of religious and classical elements, which creates a contrast between the different aspects of the composition. The presence of religious figures, such as the man in red robes and the man in brown robes, suggests a spiritual or sacred context. On the other hand, the classical elements, such as the columns and the architectural style, evoke a sense of history and tradition. This contrast between the religious and classical elements adds depth and complexity to the painting, inviting viewers to explore and interpret the scene in various ways.

In the painting, there is a contrast between the elements of the composition, which adds depth and interest to the artwork. The central figure, a man wearing a crown, is depicted in a larger scale, occupying a significant portion of the image. This emphasizes his importance and prominence in the scene. On the other hand, the two smaller figures, a child and a man, are positioned in the foreground, creating a sense of depth and perspective. The child is holding a guitar, which further adds to the composition's visual interest. The combination of these elements creates a dynamic and engaging artwork that captures the viewer's attention.



The painting showcases a striking contrast between the elements. The central figure, the pope, is depicted in a regal and serene manner, surrounded by opulent details. In contrast, the figures on either side, St John the Baptist and St Paul, are portrayed with a more rugged and earthy appearance. The cherub playing the musical instrument adds a touch of innocence and lightness to the composition. The architectural elements, such as the columns and arches, further emphasize the grandeur and solemnity of the central figure.

The painting employs a striking contrast between the ethereal, heavenly figures in the sky and the earthly, detailed figures below. The sky is a deep blue with luminous white clouds, while the ground is depicted in rich, warm colors. This contrast draws the viewer's attention upwards, emphasizing the divine nature of the central figure and the significance of the event.



The painting presents a rich contrast between the celestial and terrestrial elements, which is a common theme in religious art. The upper portion of the composition features a group of figures in a state of ascension or divine intervention, with one figure prominently in the center, arms outstretched, seemingly ascending into the heavens. This figure is surrounded by other figures who appear to be in a state of rapture or divine revelation, with their hands raised and faces expressing a range of emotions from awe to ecstasy. The contrast is further emphasized by the bright, ethereal light that emanates from the central figure, which seems to be a divine source, illuminating the surrounding figures and creating a sense of divine presence. This light source is a visual cue that suggests a divine intervention or a moment of spiritual transformation. Below this group, the lower portion of the painting depicts a more grounded, earthly scene.

[Previous](#)
|
[Next](#)

Figure 38: Example of what one tab looks like for the side-by-side comparison task. The header shows the topic of the narratives and the task. Additionally a tab number is shown to give notice to the participant how much time the survey takes. A participant can click on the text boxes, which will then light up green, showing that this narrative is favored over the other.

One participant is asked to judge the narratives on a subset of 4 topics, which are chosen uniformly at random. That is 2 topics for the side-by-side comparison part and 2 for the part where the participant is asked to highlight segments of the text. In the former, for each topic each of the possible model comparison combinations is shown. There are 3 models, which can be matched up against each other in 3 possible ways. That makes 3 comparisons per topic and so 6 comparisons in total per participant. For the latter, a narrative per model is shown. This also makes per topic 3 texts asked to be judged and in total 6. The artworks are as well chosen uniform randomly from the set. This all gives a uniform distribution over the different topics, models and artworks, adding up to a length of 20 minutes.

(2/4) Focus Point

Highlight non fitting parts in the text

Judge the quality of the **text** on the **focus point** of the paintings. Do this by highlighting parts of the texts that you think are not fitting or could be improved.



The focus point of this painting is the central figure of Jesus, who is seated on a raised platform and engaged in a conversation with a group of individuals. The positioning of Jesus, with his head slightly tilted and his body turned towards the person in front of him, draws the viewer's attention directly to him. The use of light and shadow on his face and the book in his hands further emphasizes his importance in the scene. The other figures in the painting are arranged around Jesus, creating a semi-circular composition that guides the viewer's eye towards him. The person kneeling in front of Jesus is particularly prominent, with their posture and the gesture of their hands suggesting a moment of supplication or revelation. The other figures, while not as prominently featured, are arranged in a way that suggests they are listening or responding to Jesus, reinforcing the centrality of his presence in the scene.



In the painting, the focus point is the interaction between the two main figures, a man and a woman, who are depicted in a close embrace. The man is sitting on the ground, while the woman is standing above him. The scene takes place in a room with a bench and a chair, which adds to the overall composition. The presence of these objects, along with the interaction between the two figures, creates a sense of depth and interest in the painting. The composition effectively captures the emotional connection between the two individuals, making it the central point of the artwork.



The focus point of this painting is the central area where Jesus is kneeling on the ground, surrounded by the other figures. The positioning of the figures, the direction of their gazes, and the placement of objects all lead the viewer's eye towards this central scene.

[Previous](#) [Next](#)

Figure 39: Example of what one tab looks like for the highlight task. The header shows the topic of the narratives and the task. Additionally a tab number is shown to give notice to the participant how much time the survey takes. By selecting the text in the boxes the participant can highlight it with a red color, indicating that it does not fit or could be improved.

The questions divided over separate tabs, one topic of composition per tab. The task of side-by-side comparisons and highlighting were alternated, one task per tab. Each tab showed the three questions as discussed above. Figures 38 and 39 illustrate examples of what the tabs look like.

In early versions the highlight tabs had the same topic and images as the side-by-side comparison tab that goes before it. Multiple participants interpreted this tab as if it was an acknowledgement of their choices in the previous tab. To make it clear that this tab has its own separate task, in later versions this also got its own unique topic and images. Additionally, when no texts were highlighted a one-time error message would pop up to give a second notice to the task description.

In the background the interface keeps track of how much time someone takes to finish each tab, to ensure the responses were serious. We disregarded responses that were done in an unreasonable amount of time.

The final version of the survey can be found by following the link¹¹.

4.4 Distribution

Mediums: In order to get people to participate in the survey, we used numerous ways to attract attention.

- *Art teachers.* Firstly we tried to get two art teachers to give the survey to their students as an assignment. During the period the students had many important projects, which left them no time for other mandatory tasks. We also tried to contact another teacher who give AI art as a subject for his lectures. Additionally we spoke to someone who worked with AI art and had connections to teachers at the art academy. Also these gave no results.
- *Posters.* We designed a poster that showed a qr-code that takes you directly to the website of the survey. This poster was put up at the art academy, multiple university buildings in Groningen, plus library buildings of Groningen and Utrecht.
- *Instagram.* We contacted many moderators of fine-art channels on Instagram, to ask if they would help distribute the message among their audience. There was no response to this question.
- *Inner circle.* We asked people we knew to help. There were multiple from this group that filled in the survey.
- *Art students.* The most effective method was to sit at the art academy and ask students personally if they would join. This made many art students, including from fine-arts, to participate. On top of this we went to the students presentations of their final projects throughout the city of Groningen. Here we talked to them about our project and discussed the survey.

¹¹<https://photosynthesismembrane.github.io/survey>

Audience: The interface included questions for the participant to fill in his or her experience in the art field. That made it possible to keep track how many people have an educated judgement on our questions. There were 12 art students and 7 attendants that either took courses in art or work in design related jobs. In total there were 33 who participated the survey.

Problems: The interface first did not work for mobile devices. This was possibly the reason why there was no response from the Instagram moderators as well as for the posters. That made us rework the interface to work on these devices as well. This gave a significant boost in participation.

For the first halve of the attendants there was no complete data collected for the side-by-side comparison. This data did not include which VLM's were matching up against each other, only the choice of the attendant was submitted. Our statistical method required this knowledge, which was lost hereby.

There was data lost for a few mobile device participants. There was a bug that did not submit the choices for them on the side-by-side comparisons. This bug was reworked shortly after.

4.5 Analysis

Side-by-side comparison: To perform statistics on this part we use a metric that is often used in online gaming, but could be used in any situation where opponents match against each other. In our situation we match VLM's against each other. This metric computes a ranking for every VLM. Each model starts with a default ranking, which changes on every match. When one narrative is chosen over another, the ranking changes accordingly for both. This ranking has a mean and variance and is assumed to be Gaussian distributed.

In the end we perform an analysis of variance on the resulting ranking Gaussians to compute whether the models are significantly differently judged by the survey attendants.

We look at all the data together to find out what model is favored overall. Then we look in more detail what model is favored for each specific topic. Also we find out whether the results are significantly different for both datasets, renaissance and pinterest.

Highlighting: To analyse the parts of the narratives that were highlighted by the participants, we perform a quantitative and qualitative measurement.

- On the one hand we calculate the proportion in the text that was highlighted. This is done by counting the number of characters that were highlighted and dividing this by the number of characters in the text. For every narrative that is done separately. Putting all the narratives together it creates a distribution of percentages. For this list of percentages we can calculate the mean and variance. We assume this distribution to be a Gaussian.

Again we perform an analysis of variance on these distributions in order to find out whether there are significant differences between sets. We look globally at the proportions being judged 'not-fitting' or 'could be improved' in the narratives of the three VLM's. This gives us insight into whether one VLM makes more mistakes than another. Also we do the same measurement in every topic separately. Additionally we look at significant differences between the datasets.

- On the other hand we do a qualitative measure by going through the results one by one and categorizing the kind of texts that were highlighted. We find common mistakes and subjects that need improvement.

4.6 Results

Side-by-side comparison: There were 33 participants in the survey. They each were shown 6 questions in the side-by-side comparison section. That is 4 comparisons for each VLM. There was some data lost, we ended up with data for 60 comparisons for each VLM.

Table 1 shows a table in which all the results are summarized. It shows how many times a VLM won over the number of matches that it took part in. The TrueSkill ratings computed for the VLM's are shown in this table as well.

Task	Llava wins	Cogvilm wins	Deepseek wins	Llava rating	Cogvilm rating	Deepseek rating
all	24 / 60	28 / 60	38 / 60	25.25 ± 1.08	26.36 ± 1.07	27.93 ± 1.08
composition	5 / 10	5 / 10	5 / 10	26.05 ± 2.92	27.90 ± 2.91	24.60 ± 2.97
balance	5 / 14	8 / 14	8 / 14	24.91 ± 2.48	25.75 ± 2.40	26.86 ± 2.40
eye	6 / 10	2 / 10	7 / 10	26.95 ± 3.12	23.07 ± 3.30	28.92 ± 3.09
movement	0 / 0	0 / 0	0 / 0	25.00 ± 8.33	25.00 ± 8.33	25.00 ± 8.33
contrast	2 / 6	4 / 6	3 / 6	18.87 ± 4.07	28.02 ± 4.03	25.45 ± 3.97
foreground	0 / 8	5 / 8	7 / 8	14.92 ± 4.78	26.74 ± 4.00	34.35 ± 4.09
symmetry	5 / 8	3 / 8	4 / 8	27.07 ± 3.45	27.05 ± 3.56	23.46 ± 3.46
focus	0 / 0	0 / 0	0 / 0	25.00 ± 8.33	25.00 ± 8.33	25.00 ± 8.33
proportion	1 / 4	1 / 4	4 / 4	20.08 ± 5.11	22.20 ± 5.14	34.43 ± 5.43

Table 1: This table shows for each VLM the number of wins over total matches along with the TrueSkill rating. The values are show overall as well as for the separate categories.

A TrueSkill rating is computed for every match that a VLM competed in. It starts at the default value of 25. All the ratings for one VLM form are assumed to be Gaussian distributed. A mean and variance is computed for the ratings for each VLM. The ratings are computed overall as well as for every subtopic of composition separately. The resulting Guassian distributed ratings are plotted in Figure 40 as well as in Figure 41.

What the analysis of variance shows is that there is a significant difference how the models are favored for the task of describing the composition in a painting. When looking at the set of all topics together we see that the VLM Deepseek is favored over Cogvilm and Llava for the particular model versions that were used. The models Cogvilm and Llava are judged to be of similar quality for this task.

When looking on every topic of composition separately in many cases we also see a significant difference in how the models are judged by the participants of the survey. The analysis of variance often shows a p-value of less than 0.05, and therefore those distributions are significantly different. This is the case for the topics 'balance', 'eye movement', 'foreground & background', 'symmetry & asymmetry' and 'proportions'. What model is favored is not always similar to the overall setting. For example, when we look at the subtopic 'how a viewer's eye moves across the painting', here the Llava model is favored over the others. There was no data gathered in the survey on the topics 'movement' and 'focus point'.

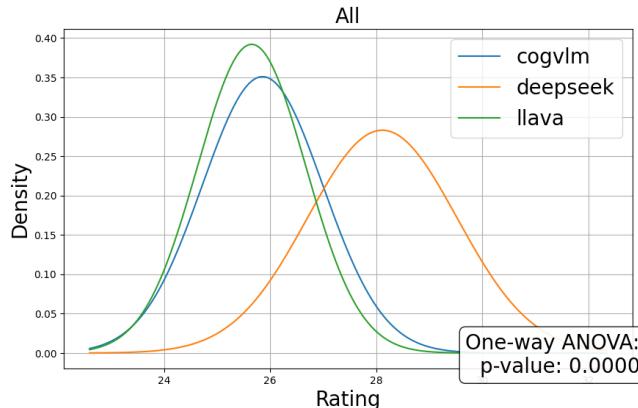


Figure 40: This plot visualizes the TrueSkill ratings for every VLM overall. A rating is a Gaussian distribution with a mean and variance. The horizontal axis visualizes the rating. An analysis of variance is performed to test whether the distributions are significantly different. The p-value is much smaller than 0.05 and so there is a difference. The VLM Deepseek has a better rating than Cogvilm and Llava.

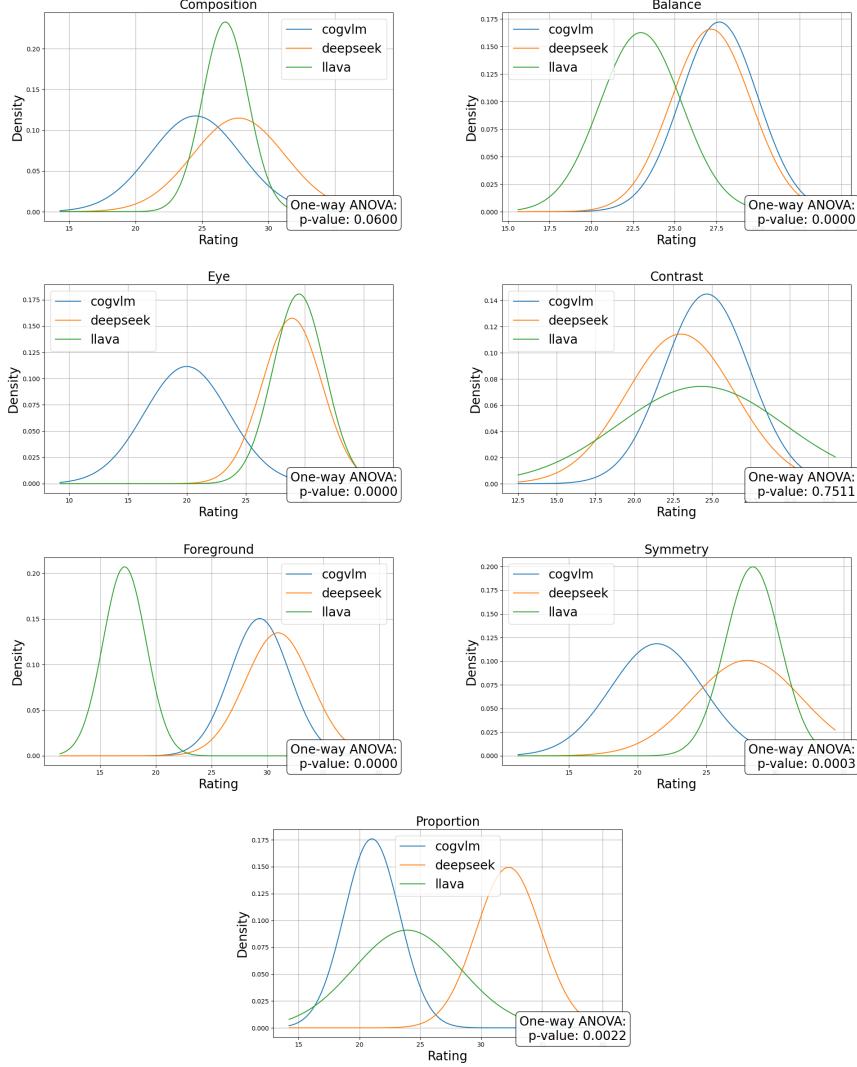


Figure 41: This plot visualizes the TrueSkill ratings for every VLM in on every separate topic of composition. A rating is a Gaussian distribution with a mean and variance. The horizontal axis visualizes the rating. An analysis of variance is performed to test whether the distributions are significantly different. The p-value of this test is shown in the bottom right, a value lower than 0.05 indicates that the difference is significant. There was no data for the subjects movement and focus.

Highlighting: From the 33 participants we collected 186 AI narratives that were highlighted on what parts did not fit or could be improved. We analysed these results quantitatively and qualitatively.

- *Quantitative evaluation.* We computed the proportion in the texts that was judged to be non-fitting or could be improved. That was done for the sets of narratives of each VLM separately. The proportions are measured in the form of a distribution. A proportion per narrative is calculated. All the resulting proportions are put in a set. This forms a distribution of proportions, with a mean and variance. The distributions from the VLM's separately, we compare to each other, whether they differ significantly. From this we can see whether the proportion of text that does not fit or needs improvement differs in the narratives from separate VLM's.

Table 2 shows the overall proportion highlighted texts for the narratives by each VLM's. Additionally this proportion is shown for each subtopic of composition.

Figure 42 illustrates how the proportions are distributed among all the narratives as described above. The distributions for narratives from each VLM individually are shown. An analysis of variance was performed on these distributions, to calculate whether they differ significantly. The p-value of 0.4219 tells us that we cannot conclude such a difference. In the VLM's art reviews on topics of composition the proportion of narrative that does not fit or could be improved is similar for the three particular model versions that we used in the survey.

Figure 43 visualizes the distributions per subtopic of composition. For each category an analysis of variance is done, from which we can almost never conclude a significant difference. Only the category 'how the viewers eye moves along the painting', the proportions of highlighted narrative differ. The narratives of the VLM Llava here shows the largest proportion of text that does not fit or could be improved. In the category 'movement' we did not gather data from the survey. The 186 gathered results are divided over 8 categories and came from 3 different VLM's, meaning that there are 27 groups of data points each containing 6.88 results on average.

Task	Llava	Cogvlm	Deepseek
All	17.98%	11.69%	13.41%
composition	20.46%	10.15%	12.06%
balance	19.5%	10.04%	7.41%
movement	0%	0%	0%
focus	8.83%	13.96%	31.2%
contrast	23.29%	4.91%	11.07%
proportion	12.39%	1.59%	5.67%
foreground	10.07%	12.37%	9.45%
symmetry	19.35%	23.36%	15.17%
eye	22.73%	12.29%	9.11%

Table 2: Percentage of text that was highlighted by participants in the survey for the various topics, showing the data for each VLM side by side.

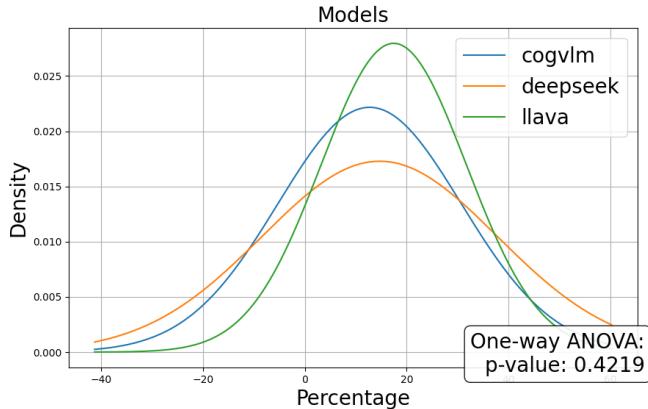


Figure 42: These plots visualize the overall results of the highlight part in the survey. For all texts was calculated what percentage was highlighted. These together form a distribution with a mean and variance. Per originating VLM the corresponding distribution is plotted. An analysis of variance shows no significant difference between the three distributions.

- *Qualitative evaluation.* All the text highlighted by the survey participants were analysed qualitatively. On these analytics we based a list of categories which characterizes common mistakes by the VLM’s. The list comprises the following categories:

(a) Nouns that are not in the picture

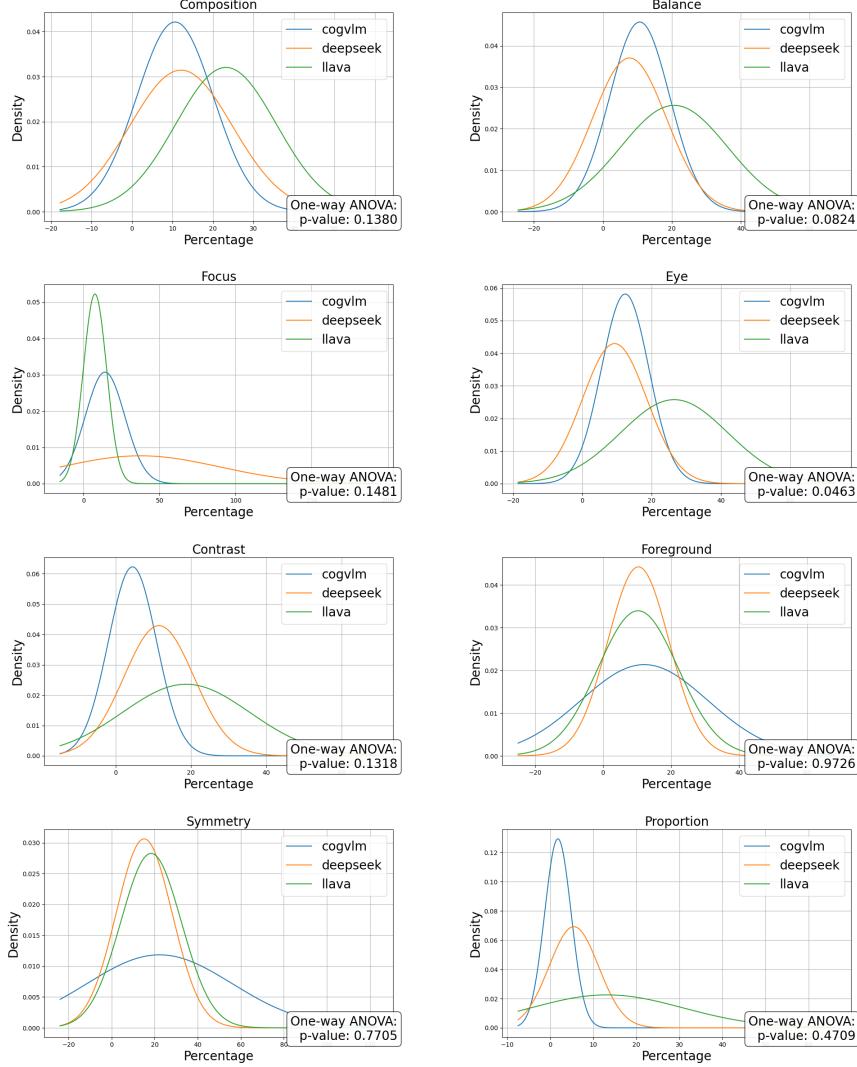


Figure 43: These plots visualize results of the highlight part in the survey. Each picture shows the results on one topic. For all texts within this topic was calculated what percentage was highlighted. These together form a distribution with a mean and variance. Per originating VLM the corresponding distribution is plotted. We did an analysis of variance on the distributions in each topic, showing that only in the topic 'how the viewers eye moves across an artwork' the highlighted proportion significantly differs per VLM.

- (b) Elements in the picture that are wrongly labeled by a noun
- (c) Not fitting adjectives

- (d) Sentences that do not describe the specific subject of composition
- (e) Odd word usage
- (f) No understanding of the concept used in art theory



Figure 44: The woman is holding an umbrella, "which adds a sense of movement and fluidity to the scene".

The following text shows examples of what a mistake in such a category looks like.

The narrative for the Figure 44 contained "The woman is holding an umbrella, which adds a sense of movement and fluidity to the scene", where the part after the comma was highlighted by a participant of the survey. This falls in the category of not fitting adjectives, since the words "movement" and "fluidity" are out of place in the context of the woman holding an umbrella. Although the couple seem to be moving, the dress catches wind, which are two elements that depict a form of movement.



Figure 45: The composition of this work is "well-balanced" and "harmonious".

Also, the ground appears watery, and the paint used resembles watercolor, creating an effect that can be described as fluid. Since these elements are there in the picture, this can make the AI more likely to use these kind of words. When they are used in combination with umbrella, they are really out of place.

For the Figure 45 the words "well-balanced" and "harmonious" were used although the scene seems rather daunting. The crows and the purplish dark colors of the trees that have these jagged branches makes it quite the opposite of harmonious or balanced. This makes this highlight fall in the category 'not fitting adjectives'.



Figure 46: The blue eye stands out prominently "against the yellow background".



Figure 47: The painting's composition effectively captures the essence of the couple's relationship "and the beauty of their surroundings".

You can see from the example in Figure 46 that there is no understanding of the concept background. The AI does not know that the face is not a background of the eye. In art theory a relation between foreground and background can be described, which are concepts not understood by the AI in this example.

An example where a part was highlighted that gives a description of the painting that is not about the topic is shown in Figure 47. The text should have been about the topic what the focus point is in the picture. The part that was highlighted was: "and the beauty of their surroundings".

The motivation of highlighting this part could have been that it is not about the focus point, since it does give a good description on the image in general.

5 Experimental setup

In order to test how well the VLM’s can be used to generate training data that encompasses analytics of art composition, we synthesize several datasets and use those to retrain an open source diffusion model. We then evaluate the resulting new diffusion models on how well they generate images in line with envisioned compositions.

Datasets: We select three differently themed sets of artworks, each of which contain 2500 images taken from the WikiArt dataset [41].

- *Renaissance*. As discussed before the renaissance period is known for its development art composition. We use such a set of artworks to capture all kinds of compositional patterns from this era to be learned by the diffusion model. For this set we took artworks from the folders ‘High_Renaissance’, ‘Mannerism_Late_Renaissance’ and ‘Northern_Renaissance’.
- *Abstract*. We compose another set of artworks from the folders ‘Abstract_Expressionism’, ‘Action_painting’, ‘Color_Field_Painting’, ‘Fauvism’ and ‘Pop_Art’. Because of the more abstract nature of these works they contain different compositional features.
- *Landscape*. Another set is composed by running a k-means clustering algorithm¹² on features captured by the Inception network in the last layers of the network, for each of the artworks. We order the whole WikiArt dataset in 50 clusters, each containing similar works. 9 clusters are selected containing landscape themed works.

Each of these three datasets we filter to select the 2500 works based on the height width dimensions. The diffusion model requires square images as training examples. Therefore images that are not of such dimensions need to be scaled and or cropped. Such modifications destroy the composition to some extend and therefore we want to do such alterations as least as possible. This is why we select the works that are the most square. All images are scaled and cropped to have a size of 512x512 to work as input for the diffusion model.

For each of these image datasets we generated composition analytics using VLM’s. The same vision language models are used as were evaluated in the survey. For each image we generated analytics on one topic of composition. All the topics were evenly distributed over the whole set of images. We created a set of captions per VLM. Additionally we created a set that contains analytics from all the VLM’s. We looked at the survey side-by-side comparison results

¹²https://github.com/TomEijkenkamp/image_similarity_clustering

for what VLM performs best in each separate topic of composition. This last dataset has analytics for each image on a topic, using the VLM that was judged best for that topic in the survey. Based on the plots in Figure 41, the VLM that received the highest rating was utilized for its respective topic.

As a result we generate four different captions for the images in each dataset:

- *Llava*.
- *Cogvlm*.
- *Deepseek*.
- *Mix*.

Baseline approaches: We compare the diffusion models we create to a baseline model stable diffusion v1.5 [35].

Training Procedure Stable Diffusion v1-5 is a latent diffusion model which combines an autoencoder with a diffusion model that is trained in the latent space of the autoencoder. During training,

- Images are encoded through an encoder, which turns images into latent representations. The autoencoder uses a relative downsampling factor of 8 and maps images of shape $H \times W \times 3$ to latents of shape $H/f \times W/f \times 4$.
- Text prompts are encoded through a ViT-L/14 text-encoder.
- The non-pooled output of the text encoder is fed into the UNet backbone of the latent diffusion model via cross-attention.
- The loss is a reconstruction objective between the noise that was added to the latent and the prediction made by the UNet.

stable-diffusion-v1-1: 237,000 steps at resolution 256x256 on laion2B-en. 194,000 steps at resolution 512x512 on laion-high-resolution (170M examples from LAION-5B with resolution = 1024x1024).

stable-diffusion-v1-2: Resumed from stable-diffusion-v1-1. 515,000 steps at resolution 512x512 on "laion-improved-aesthetics" (a subset of laion2B-en, filtered to images with an original size $\geq 512 \times 512$, estimated aesthetics score > 5.0 , and an estimated watermark probability < 0.5 . The watermark estimate is from the LAION-5B metadata, the aesthetics score is estimated using an improved aesthetics estimator).

stable-diffusion-v1-5: Resumed from stable-diffusion-v1-2 - 595,000 steps at resolution 512x512 on "laion-aesthetics v2 5+" and 10 % dropping of the text-conditioning to improve classifier-free guidance sampling.

Validation metric: We generate images using prompts that ask for certain

compositions. We perform a quantitative measure by calculating the Clip score on these generations. According to the study of OpenAI [3] when training a system with long detailed texts, this system also requires such detail in the prompts. Therefore we test the diffusion models with prompts generated by the same VLM’s. We select a validation set containing 40 works for each of the dataset themes, renaissance, abstract and landscape. The works we select were not used for training. For these validation images, we generate analytics for all the 9 topics of composition. One prompt is generated per VLM. Combining this creates $40 \times 9 \times 3 = 1080$ prompts. For each of the 12 models that we trained and the stable diffusion v1.5 model we generated 1 image per one of these prompts.

Additionally we do a qualitative study on the results. All the generated images were laid side by side for each of the models in an interface along with the prompt that was used to generate that image.

Implementation details: We synthesized captions for the datasets using code provided by the developers of Llava¹³, Cogvlm¹⁴ and Deepseek¹⁵ respectively. This code is combined¹⁶ to create a smooth pipeline that creates all captions in one run.

To retrain the open source diffusion model on our synthesized datasets the code from the Diffusers Github repository¹⁷ was used. Our additions can be seen in our fork¹⁸ of the original repository. This contains a complete set up to retrain any diffusion model on any selected dataset. Hyper parameters such as number of iterations and learning rate can be refined to suit the situation.

The repository shows an example where they train stable-diffusion-v1.4 with a naruto dataset consisting of about 1400 images. For our models we take a similar approach, using the same hyper parameters, only changing the number of iterations. The open source diffusion model we use for retraining is the same as our baseline: stable-diffusion-v1.5.

We did four runs of training our models, altering the number of training iterations from 8k, 15k, 25k to 35k.

To generate evaluation prompts we used the same repository¹³ which was used to synthesize captions for datasets. To generate evaluation images a simple script was used which takes a specific diffusion model and set of evaluation prompts to generate from¹⁹. This repository contains code to compute clip scores on the image prompt combinations. Additionally it consist of a HTML interface, which shows all the generated images in categories and sets results from DM’s side by side. We generated almost 60k images and because of that an online web interface on Github did not work.

¹³https://huggingface.co/docs/transformers/model_doc/llava

¹⁴<https://huggingface.co/THUDM/cogvlm-chat-hf>

¹⁵<https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat>

¹⁶https://github.com/photosynthesismembrane/caption_generator

¹⁷https://github.com/huggingface/diffusers/tree/main/examples/text_to_image

¹⁸https://github.com/photosynthesismembrane/diffusers/tree/main/examples/text_to_image

¹⁹https://github.com/photosynthesismembrane/retraining_results

6 Results

Quantitative evaluation: We calculated the distributions of Clip scores for all the generated images from the diffusion models, plotted in the Figures 48 to 51. The generated images from the diffusion models trained on the different datasets, renaissance, abstract and landscape, are combined. We show the distributions of Clip scores for the diffusion models trained with captions per VLM. This results in four distributions of Clip scores from our own models, trained with Llava, Cogvilm, Deepseek and a mixture of VLM captions. Additionally the baseline Clip scores are shown from the stable diffusion v1.5 model generations. Results from the diffusion models trained with captions combining all VLM's in the plots is named 'best'.

We used analysis of variance to compute whether the distributions are significantly different, showing that in all cases they are.

The diffusion models trained for 8k, 25k, 35k iterations all show the same order in performance. The models that use a mixture of captions based on best qualified VLM's on composition topics, comes out with the best Clip scores. Then the DM's trained with captions from Cogvilm and Deepseek are performing similarly, but lower then the 'best' diffusion model. The DM trained with Llava captions has lower Clip scores.

The diffusion models trained for 15k iterations show a similar order, only the 'best' DM's are here scoring lower then the Cogvilm and Deepseek variants.

The 'best' diffusion models perform best when trained for 25k iterations. All the other models perform best when trained for 15k iterations. Still the same order in performance comparison applies.

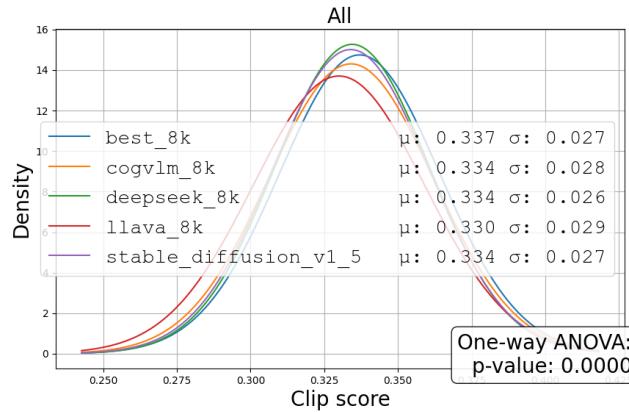


Figure 48: Clip scores for diffusion models trained in 8k iterations.

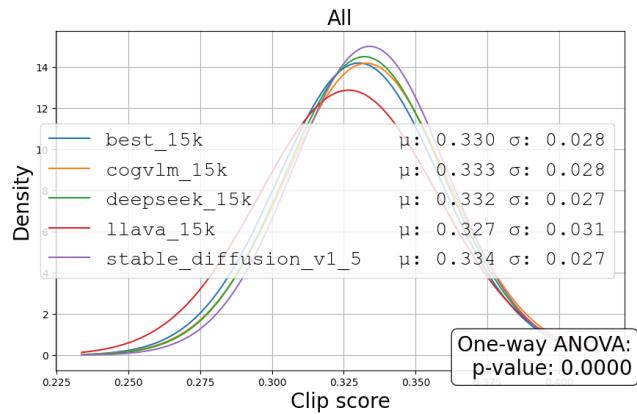


Figure 49: Clip scores for diffusion models trained in 15k iterations.

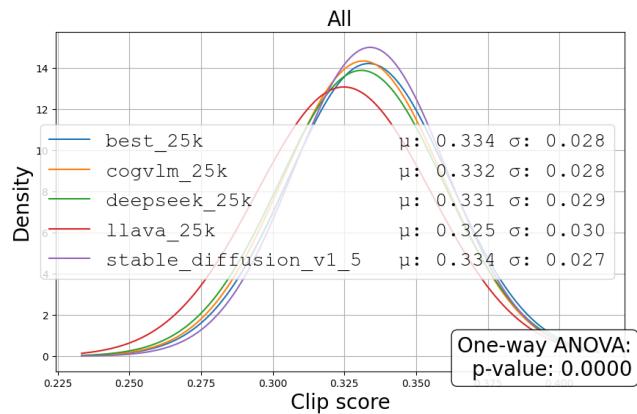


Figure 50: Clip scores for diffusion models trained in 25k iterations.

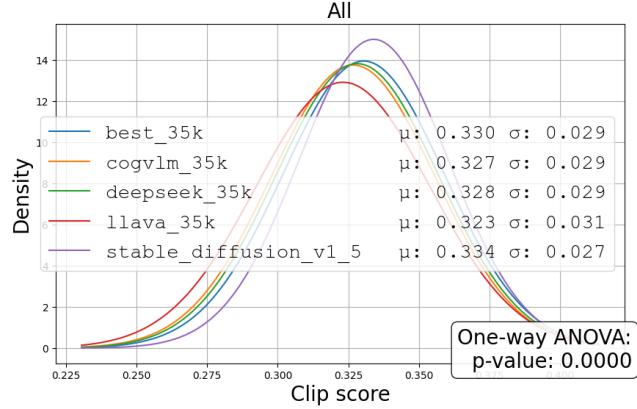


Figure 51: Clip scores for diffusion models trained in 35k iterations.

Figure 52 shows the loss over time for the diffusion model trained with renaissance dataset with llava captions over 25k iterations. Here you can see that the loss still decreases over that time. The validation loss goes up at 35k iterations, indicating that further training does not improve results.



Figure 52: Loss diffusion model trained with renaissance dataset with llava captions over 25k iterations.

We still see that the baseline model stable diffusion v1.5 scores best in most cases. This model is trained in an entirely different fashion, which makes this comparison not fair. It might be that retraining this model for 8k in our scenario showcases the best set up, since there our diffusion models show similar results or perform better than the baseline. Retraining a diffusion model like we

do is very sensitive to hyper parameter settings and vulnerable for catastrophic forgetting, and so requires precise fine tuning to create well performing diffusion models.

Qualitative evaluation: From qualitative evaluation you can see that the topics of composition are quite well learned. In Figure 53 you see that the focus point is described in the prompt is visualized very well in with the retrained models.



Figure 53: A sample of generated images that show a focus point, from respective models. Generated from the prompt: "In the painting, the focus point is the white circle, which is surrounded by various shapes and colors. The white circle serves as a central element in the composition, drawing the viewer's attention and creating a sense of balance and harmony within the artwork. The surrounding shapes and colors add depth and complexity to the painting, making it an interesting and engaging piece of art."

In the sample shown in Figure 54 shows that the best model accurately depicts many elements in the prompt. The Coglm model also shows the two figures in their requested poses. The order of the models is the same as the ones from earlier.

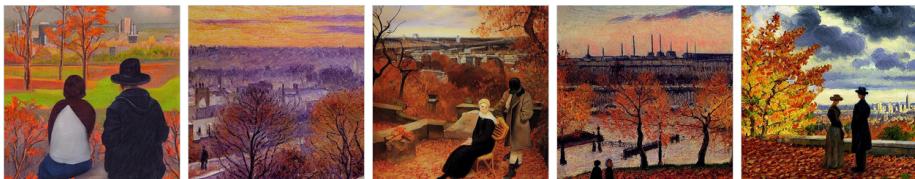


Figure 54: Many of the elements from the prompt align to what is seen in the generated images, from respective models. Generated from the prompt: "The painting showcases a beautiful contrast between the elements. The warm hues of the autumn leaves contrast with the cooler tones of the sky and the distant cityscape. The two figures, one seated and one standing, add another layer of contrast, with their dark attire juxtaposed against the vibrant backdrop. This balance creates a harmonious visual experience, drawing the viewer's attention to various parts of the canvas while ensuring that no single element overshadows the others."

A sample from the renaissance models show proportion, background and a general well interpretation of the prompt in the retrained models. Figure 55 shows the generated images, the order is the same as above. The prompt is show in the caption.



Figure 55: A sample of generated images from respective models that show well aligning concepts: proportion, background and other details from the prompt. Generated from the prompt: "The painting depicts a figure, presumably a man, engrossed in reading a large open book. The figure is dressed in a robe, and the book's size is significant, covering a large portion of the figure's lap and extending outward. The background is dark and nondescript, which creates a stark contrast with the figure and the book, making them the focal points of the composition. The contrast is further emphasized by the lighting in the painting. The light source appears to be coming from the upper left side, casting a soft glow on the figure's face and the book, while the rest of the scene remains in shadow. This directional lighting creates a sense of depth and draws the viewer's eye towards the illuminated areas, enhancing the focus on the figure and the book. The figure's posture, with one hand resting on the book and the other supporting it, suggests a moment of contemplation or study."

Often it is not so easy to notice the difference between the models on their knowledge of the topics of composition. In Figure 56 you see that all models generate well aligned pictures.



Figure 56: General well performance of models on the prompt: "The viewer's eye movement is guided through the painting by the pathway that leads from the foreground to the church in the background. The pathway acts as a leading line, drawing the viewer's gaze deeper into the scene and towards the focal point of the church."

Stable diffusion quite regularly makes a border surrounding the painting, while the retrained models do not. They are trained on pictures that exactly capture the painting and so this might reduce the chance that a border is made in generation. Example is show in Figure 57.



Figure 57: Stable diffusion often displays a painting border in the images. Results generated with the prompt: "In the painting, the focus point is the group of people sitting on a bench in the foreground. The scene is set in a village, with a woman sitting on a bench next to a building. The presence of the bench and the people sitting on it adds a sense of depth and interest to the composition. The other elements in the painting, such as the dog, the chicken, and the other people, contribute to the overall atmosphere and storytelling of the scene. The combination of these elements creates a visually engaging and dynamic composition that captures the viewer's attention."

7 Discussion

In order to give diffusion models knowledge on effects that can be used by positioning elements within an artwork or graphic design, this knowledge can be inserted into the training data. The model can be trained with image text combinations, where the text describes compositional effects that are used in the corresponding image.

Such training data can be created automatically using vision models. VLM's with a chat function are versatile to analyse images on many topics, because of their natural language interface. Such models can be prompted to describe any kind of aspect of an image and therefore they can be used to generate texts that describe all kinds of compositional effects that can be seen in images. Specific prompt engineering is required in order to get the right data.

In our survey we find that people favor one VLM over another on specific art composition analysis tasks. We asked the people to evaluate responses from the VLM's Llava, CogvIm and Deepseek. The order in which the VLM's rank differs per task. It is the overall quality of the VLM's response by which a VLM is favored. The quantity of valued text within a response is similar among the VLM's. We evaluated what participants considered not fitting in narratives from the VLM's and categorized common mistakes.

It pays off to first analyse what VLM is best to use to analyse particular patterns in images. In our experiment we combine analytics on various topics as

training data for diffusion models. When utilizing the VLM that performs best on each task specifically and combining the results, we created a better model in comparison to using one VLM for all the tasks.

Further analysis of the highlighted texts in the VLM's narratives on composition could be done. An art expert could be used to do such a review for a more qualified opinion. Additionally large language models can be used to categorize the highlights.

Although Deepseek-vl and Cogvilm were tested with their latest models, there is a newer model for Llava: llava-hf/llava-v1.6-mistral-7b-hf. This possibly has improved results on the tasks and there explain the shortcomings compared to Deepseek and Cogvilm in this study.

The VLM's are trained on different datasets that overlap to some extend. Further research could investigate training a VLM with significant datasets of art analytics and theories, history of art movements. This might have a established effect on the quality of analysis of artworks.

The inner architecture varies as well for the VLM's. Cogvilm claims to learn a deeper patterns between visual and language relations, in which they differ from more surface level learning VLM's such as Minigpt. Such deeper patterns might be required to analyse intricate concepts in art such as well designed compositions.

Diffusion models are said to generate an average over the training data. This is different from generating an artistic composition, since the training data might consist of many examples lacking an art design. When the training data consist of much knowledge about composition analysis, the model might learn a larger skew in these kind of patterns, therefore being more prone to use these kind of patterns when generating images. Still there is no hierarchy learned in what is better to use in what situation. Future research could study the process of design in terms of decision making, why diffusion models generate certain graphic designs over others, and how to make them output quality aesthetic.

How in the human brain language is connected to visual thought is not simple [39]. An artist undergoes a whole process going from goal to end result. Throughout the life he or she gained knowledge and practiced the skill. Additionally humans have emotions and feelings. Future research could further investigate the learned mapping between language and visuals in diffusion models and thereby inspire from human behavior.

Although the text part of the training data is improved in detail on composition, the images are not required to be of compositional quality. Additional research could look at how to include examples of compositions in all kinds of categories.

During the extend of the project there were many comments of artists about the fact that a diffusion model uses artworks as inspiration to generate new pieces. There are cases where this was done without the consent of creators of the works. Computers are said to work different from a human brain, computers have exact memory which is not the case for humans, therefore they always copy. Output is pointed out containing clear styles used by real artists. Many artist were against AI art research or not willing to contribute because of such

reasons. Future research in whether these systems are copying styles of artists and therefore stealing their work is important. Systems that build their own style or check for plagiarism is necessary.

As said in OpenAI's research, the system requires prompts of similar detail as in the training data. They used an LLM to upscale the prompt to such detail. Such an upscale was not studied in this thesis.

The results here are compared to Stable Diffusion which is a model trained from scratch. This uses 500 million training examples. We retrain this model with 2500 images. This is very vulnerable for catastrophic forgetting. This setup is not comparable to training a system from scratch. Future work could comprise training from scratch for fair comparison.

We tested generated images with the Clip model on similarity between an image and prompt embedding. This Clip model might not capture compositional aspects on quality level. This model is not specifically trained on artworks and art studies and therefore might not pay as much detail to artistic patterns. An art expert model described above could do better analysis of art patterns. Such a hypothetical model could be a better measure when used similarly to calculate similarity between prompt embeddings and embeddings of the related generated images.

In our experiments we generated images which were only tested qualitatively and quantitatively. Additionally generated images should be analysed in a human evaluation, similar to the survey set up for analysing the VLM's art knowledge.

The validation of the trained diffusion models was only done at the end of training. To gain better insight into the performance, overfitting or catastrophic forgetting, testing and or validation during training must be done. As said this type of training for these models is vulnerable and therefore it takes sensitive tuning to make them perform right.

8 Conclusion

In this thesis we showed that we can integrate diffusion models with artistic skills by training them with data that includes analytics on artwork composition. VLM's are very versatile and can be used to generate narratives on all kinds of topics around composition. These models differ in their expertise and so searching for the right model for the task is rewarding.

Acknowledgments

Special thanks to artist Bager Akbay for support in selecting the thesis topic, and to artist Evgim Kavcar for her guidance in structuring the survey.

References

- [1] W. Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.
- [2] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.
- [3] J. Betker, G. Goh, L. Jing, TimBrooks, J. Wang, L. Li, LongOuyang, JunTangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and A. Ramesh. Improving image generation with better captions.
- [4] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.
- [5] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [6] DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [7] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. Construction and Analysis of a Large Scale Image Ontology. Vision Sciences Society, 2009.
- [8] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis, 2021.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners, 2021.
- [11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn, 2018.
- [12] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.
- [13] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

- [14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020.
- [15] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022.
- [16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.
- [17] T. Inglis. *The Graphic Design Bible: The definitive guide to contemporary and historical graphic design*. Octopus, 2023.
- [18] G. Kim, T. Kwon, and J. C. Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022.
- [19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything, 2023.
- [20] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [21] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.
- [22] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language, 2019.
- [23] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.
- [24] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023.
- [25] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.
- [26] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, and C. Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.
- [27] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022.
- [28] J. Mao, X. Wang, and K. Aizawa. Guided image synthesis via initial image editing in diffusion model, 2023.
- [29] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

- [30] E. A. Payne. *Composition of Outdoor Painting*. DeRu's Fine Arts, Bellflower, California, 1st edition, 1941.
- [31] W. Peebles and S. Xie. Scalable diffusion models with transformers, 2023.
- [32] 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.
- [33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection, 2016.
- [34] A. Riechers. *The Elements of Visual Grammar: A Designer's Guide for Writers, Scholars, and Professionals*. Princeton University Press, 2024.
- [35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [37] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.
- [38] N. Shazeer. Glu variants improve transformer, 2020.
- [39] G. Sullivan. *Art Practice as Research: Inquiry in Visual Arts*. Sage Publications, Thousand Oaks, Calif., 2nd edition edition, 2010. Includes illustrations; 23 cm.
- [40] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. Eva-clip: Improved training techniques for clip at scale, 2023.
- [41] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka. Improved artgan for conditional synthesis of natural image and artwork, 2018.
- [42] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.
- [43] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.
- [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

- [45] A. Vidhya. Automated image annotation using auto-annotate tool. <https://medium.com/analytics-vidhya/automated-image-annotation-using-auto-annotate-tool-f8fff8ea4900>, 2024. Accessed: 2024-08-17.
- [46] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang. Cogvlm: Visual expert for pretrained language models, 2024.
- [47] C.-H. Yeh, T.-Y. Cheng, H.-Y. Hsieh, C.-E. Lin, Y. Ma, A. Markham, N. Trigoni, H. T. Kung, and Y. Chen. Gen4gen: Generative data pipeline for generative multi-concept composition, 2024.
- [48] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training, 2023.
- [49] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.