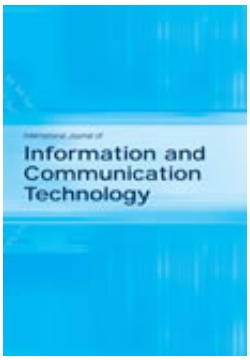


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Application of an AI-driven visual aesthetic scoring system for style calibration in art works

Feng Tan, Mei Wang

DOI: 10.1504/IJICT.2026.10076004

Article History:

Received:	07 October 2025
Last revised:	24 November 2025
Accepted:	30 November 2025
Published online:	11 February 2026

Application of an AI-driven visual aesthetic scoring system for style calibration in art works

Feng Tan

Modern Logistics and Intelligent Manufacturing College,
Wuhu Vocational Technical University,
Wuhu, Anhui, 241003, China
Email: wuhutanfengei@163.com

Mei Wang*

Office of Scientific Research,
Wuhu Vocational Technical University,
Wuhu, Anhui, 241003, China
Email: 18955362512@163.com
*Corresponding author

Abstract: This research explores AI-driven visual aesthetic scoring systems as tools for evaluating and refining artistic styles, with a particular focus on interior design and computational modelling. The study demonstrates how artificial intelligence can enhance artistic quality and align computer-generated imagery with human aesthetic preferences. By integrating compounded loss functions, curated datasets, and diffusion-based architectures, the model significantly improves visual appeal, stylistic consistency, and task performance. A composite loss-based AI framework was developed using a customised interior design dataset annotated with style tags, aesthetic ratings, and spatial attributes. The system, fine-tuned with user-defined parameters, produced results that were both visually appealing and contextually appropriate. Experimental outcomes revealed statistically robust improvements of 52.54% in portal engagement (Cohen's $d = 1.69$, $p < 0.001$, 95% CI: [47.8%, 57.3%]) and 40.08% in agency engagement ($d = 1.52$, $p < 0.001$), validated through rigorous statistical testing including permutation tests, bootstrap resampling, and multiple comparison corrections. User studies further indicated that AI-selected or AI-generated images were preferred over other sources, receiving higher aesthetic ratings and engagement levels.

Keywords: visual aesthetic scoring; style calibration; AI in art; diffusion models; aesthetic evaluation; artistic style transfer; generative design; computational aesthetics.

Reference to this paper should be made as follows: Tan, F. and Wang, M. (2026) 'Application of an AI-driven visual aesthetic scoring system for style calibration in art works', *Int. J. Information and Communication Technology*, Vol. 27, No. 9, pp.39–69.

Biographical notes: Feng Tan is working at the Wuhu Vocational and Technical University and has a research interests include art design, animation design, and intangible cultural heritage communication.

Mei Wang is a Lecturer and has a specialising in art design, animation design, fine arts, and the dissemination of intangible cultural heritage.

1 Introduction

Things that make you feel good, like, or think are beautiful, you should care about how they look. The fact that art is a big part of psychology is an easy way to explain what it is. When people like something, like art or their city, they look at it longer. The things they look at show how interesting something is to them. Some people say that aesthetic experience is when your senses, mind, and thoughts all come together (Ishrat and Abrol, 2020). The eyes help us see, understand, and make sense of the world around us. When people look at something, it is hard because they have to figure out how the shapes, colours, textures, and space all fit together. A lot of people find something beautiful that they do not understand. When there are many parts, it can be hard to understand the whole. The word 'complexity' can be used in a number of different ways. Also, some say that the growth of Gestalt psychology (Marin and Leder, 2022) shows how important it is for our brains to be hard to understand. This is because the brain might be able to tell how hard a piece is by the way it is put together. Something that you look at helps you learn about it. The information is then put into groups and organised. The building plans of a person can show a lot about how they choose the most beautiful parts of their home and how they feel about it.

Empirical aesthetics (Beder et al., 2024) has looked at the order of building faces. A lot of research has been done on how building groups work and how people pick the shapes that look best at different speeds. But no one has yet looked into how the organisation of Gestalt principles and how hard they are to understand affect how people feel about the outside of buildings. The first question this study tried to answer was whether the way Gestalt processes information changes how people see the beauty of harder or easier building materials. Abstract, non-representational forms were used by the movement to try to show order and unity. There is a clear structure in neoclassicism that can help us understand the rule-based logic of AI programs (Cetinic and She, 2022). Shape and colour are important parts of this structure. Even though generative AI art is still growing, some people still do not agree on some things. A lot of study has been done on the technical side of AI systems, like how algorithms that learn art styles help them come up with new results. But not a lot of studies have been done on whether these AI systems can understand the rules of composition that go along with certain art styles, like neoclassicism. We have mostly looked at how well AI can copy and learn the most basic parts of different art styles so far. However, translating these theoretical aesthetic principles into computational models requires explicit formalisation. While Gestalt psychology emphasises symmetry, balance, continuity, and closure, and neoclassicism prioritises colour harmony, geometric regularity, and tonal balance, most AI systems lack mechanisms to encode these concepts mathematically. This study addresses this gap by defining eight quantifiable aesthetic descriptors – including symmetry coefficients, colour harmony metrics based on Itten's theory, and compositional balance measures – that transform abstract artistic principles into algorithmically tractable features integrated directly into model training objectives.

For example, we have looked at how colours are organised and how brush lines are made. A few experts have only looked at how AI and machine learning can be used to make abstract and modern art come alive. A lot of art has not been looked at closely enough for broad readings, like the expressionism of the abstract art trend (Zhou and Lee, 2024). On the other hand, neoclassicism is a very organised and rule-based approach that is very interested in colour harmony and geometric shapes. Neoclassicism is the name for a surface that has a vertical line, a horizontal line, and black, grey, and white that are not coloured at all. This makes it a great subject to study how well AI systems can deal with formal ideas of beauty (Mun and Choi, 2025). But there have not been many real-world studies that look at how well AI can make works that follow the neoclassical style of balance, harmony, and order. A study of this kind is very important because the neoclassical structured method should work well with computational analysis. For this reason, it is important to study this gap. Here's the first thing: AI is still very important in the visual arts. This means that artists, curators, and scholars can better rate the artistic value of AI-made works if they know how they relate to aesthetic principles.

As AI technology quickly develops, especially in computer vision, entropy, and information theory, using AI to make art has become a popular and important topic in both the public and academic worlds. In the past few years, AI drawing tools have grown in popularity, especially among people who use them often. This has led to a lot of study in the classroom and in the real world into how they can make artists more creative (Wang et al., 2024). These tools help people get past creative blocks and come up with new ideas. They also make it easy and quick to make art. AI painting tools like DALL-E, Stable Diffusion, and MidJourney use deep learning to make photos that look great. The tools can understand and look at many types of art. They can also teach you styles and techniques that you can then use in your own work. AI drawing tools are great for artists who want to try out new styles and methods. It is also easy to change the creative settings on AI drawing tools so that users can make their own unique works of art (Xu et al., 2023). Producers are very interested in how easy it is to use, which gives artists new ways to get ideas and explore their talent. It's easier for AI and art to work together because of these things. This helps art grow in new ways. More and more research is being done in this area, but there are still questions about how to test how AI drawing tools affect people's artistic creativity and what causes these effects (Tian et al., 2025).

How creative people are depends on a lot of things, like how they are different from others, what they are doing, and their own specific creative goals. When people used to judge artistic talent, they mostly relied on reviews from experts or their own opinions. They did not use structured quantitative analysis. The goal of this study is to make a more objective and accurate deep learning rating model so that we can find out how much regular people can improve their art skills when they use AI painting tools. The two problems below are the main ones that this work aims to fix.

This paper's structure is setup as follows: the related works are presented in Section 2, with an emphasis on the use of an AI-driven visual system. The suggested aesthetic scoring system is described in full in Section 3's methodology section. The experiments and findings are covered in Section 4, with a focus on style calibration in artwork. Section 5 wraps up the work by summarising the main conclusions and suggestions for further research.

1.1 Contribution of this study

A system based on composite diffusion is presented in this study, which helps the field of AI-driven aesthetics move forward. This framework not only creates interior designs that look good, but it also makes sure that designs that look good meet useful and stylistic standards. It is a big deal that the ADSSFID-49 dataset was created. It includes aesthetic scores, decoration styles, and spatial usefulness in a way that has never been done before in research. Each of these parts has not been used together in studies before. Standard ways of making images usually only focus on making them look good, but this study goes further than that. To do this, a new composite loss function is introduced that combines these three things. Using artificial intelligence to make complicated artistic ideas workable and create designs that are both nice to look at and useful in real life is what the integrative approach is all about. Artificial intelligence-generated aesthetics have been shown to work in real-world situations, which is another important addition. The study not only compares different generative models, but it also uses user-based ratings to look at how engaged people are with the models. When compared to baseline sets, the results of this review show that the perceived quality of the aesthetics is more than 50% better.

Also, using AI-selected images in real estate shows that they have a commercial effect, as big increases in viewer clicks have been seen across both portals and agencies. No matter if the images are commercial or not, this is always the case. The study creates a scalable strategy for matching the creativity of artificial intelligence with what people want and what's useful for business. This structure is setup by bridging the gap between technical innovation, stylistic alignment, and real-world use.

2 Related works

Deep learning and generative models have come a long way in the past few years. This has caused a lot of new types of art to grow. We have learned a lot about the pros and cons of AI-made pictures, as well as how to use and gain from them. We carefully look at several different generative models to see how well they follow style and time limits. This is on top of what has already been done. A study that combined convolutional neural network representations of text and style was the first to think of neural style transfer. Brain style change was possible because of these things. The new study built on these ideas, which made styles cheaper and gave people more control over how they are used (Png et al., 2024). Spread-based models have gotten better at making high-fidelity soundtracks over the last few years. This is one way that styles can change in more complicated ways. Our study is unique because it looks at a number of cutting-edge models from various art forms and time periods. It's now easier to see how well the model worked (Barros and Ai, 2024). A lot of people have tried to figure out how to look at pictures made by AI and figure out what they are.

The CIFAKE study project tries to find out what real computer-made pictures are and how to tell them apart from real art. Similarly, we looked at mean AI art to show how hard it is to both show that something is real and hide it at the same time. Do these kinds of things when you want to know if something is real or not. This is especially important when you care about how other people see things. A lot of large files (Asperti et al., 2025) can help with this kind of research:

- *ArtiFact Dataset:* A lot of different real and fake pictures are in this collection. There are portraits, landscapes, cars, works of art, and faces of people and animals. In this group, some pictures were made 25 different ways. There are 13 GAN-based models and seven diffusion models used in these ways.
- *Wild Fake Dataset:* A set of data that will be used to see how well AI-made programs for finding pictures work in different settings. Fake pictures from the open-source community make it up. They show a lot of different styles and ways to put pictures together.
- *TWIGMA Dataset:* A huge group of pictures pulled from Twitter between 2021 and 2023 and made by AI. In the collection are things like tweet content, engagement numbers, and hashtags that are linked to the images.

These tests are also meant to see how well these pictures copy art that people made. Our main goal is to find pictures that AI made. What does everyone think? This helps us figure out if the drawings could be mistaken for real art made by people. This tells us something about how well the models can trick a real person. More and more, picture sets that have already been made are being used to help computers learn and study. It is not enough to just find something. What does it mean when AI pics are used to make models for machine learning? It is talked about in this piece. A lot of things come up when people look into whether or not fake datasets can be used to make machine learning better. These include bias, truth, and what is right and wrong. In the past few years, more articles have been written about digital pictures and what they can do. AI and computational aesthetics are not usually used in art and aesthetics for business reasons (Sheng et al., 2021). It is true that they are sometimes, though. It's not being looked into as much as it used to be, on how to show that computer design methods work. Fashion works will be the main things that are talked about and looked at.

They show why you should use looks when you're doing business online. There are some tools that online clothing shoppers can use to get a sense of how nice an item looks. Brain and body tests are done on people so that a database can be made that can be used to rate how pretty pictures are. These pictures are usually taken from clothing shopping websites. Then, you should add more picture features and use the 'best-first' search method along with the envelope feature selection method to find the best set of features that will help you make the best guesses (Kanwal et al., 2021). They can use this to make a better system that can guess things that are very likely to be true. It is also about the fashion business, but that's not what the work is about. In order to give users what they want, they want clothing advice systems to know how clothes look (Fernandez et al., 2012). The Weka Machine Learning Package was used to teach and test the machine. The 500 shots were then used to teach a support vector regression model what to do. It was checked five times to make sure the model values were correct after they were trained. It works best when all of the feature sets are used, as shown by the RMSE number of 2.09. Their method is much more effective than more complex ones because it can figure out what users like about something that looks good.

2.1 AI-driven application

In the past few years, more and more places where people learn have gone digital. This makes it clear that tools and processes need to be better combined using more adaptable

and effective tech options. AI is one of the most important things to have grown in this world of constant change. One of the most important technologies in teaching and learning is code (Zhang, 2024). This is because code can be used to make algorithms that can make suggestions, predictions, decisions, and learn from different situations. Because of this rise, a new way of teaching has come about that relies heavily on data. This rise happened because more people wanted tasks to be automated, learning to be personalised, and content to be made on the spot. Because of this, many thoughts about how to use AI in school have grown in importance. Smart tutoring systems, virtual helpers, immersive and interactive experiences, and using data to get the best results are a few of these ideas (Chng et al., 2023). AI has been used in education a lot more quickly in the last ten years, thanks to tools like machine learning, natural language processing (NLP), and deep neural networks. For all of these methods to work, they need a lot of training data. It is still true, even though there's a clear way to study AI in schools.

For this reason, application programming interfaces (APIs) are very important. These let learning management systems (LMS) and other tools share data and talk to each other. It's easy to add things like delivering content, grading, real-time communication, and places where people can work together. This makes it simple and easy to integrate APIs, which helps platform interoperability without having to go through a lot of steps (Pérez-Jorge et al., 2025). A lot of people use them, especially in higher education, where they make it easy to do all of these things at once. A number of well-known software companies make APIs that can be used in schools. Some features can be used because of these APIs, such as keeping track of performance, learning data, and personalisation. Art education is currently being impacted by deep learning and AI-powered digital image processing. Now that AI has CNNs, GANs, and NST, artists have more freedom to try out different styles and effects. Innovative methods for creative learning emerge from problems like edge detection, segmentation, and super-resolution. Platforms like Deep Dream and Runway showcase AI-assisted art (Wang et al., 2025a). Many are concerned about the loss of traditional abilities, ethics, and who should be credited for the results when AI provides unique and quick feedback to enhance learning.

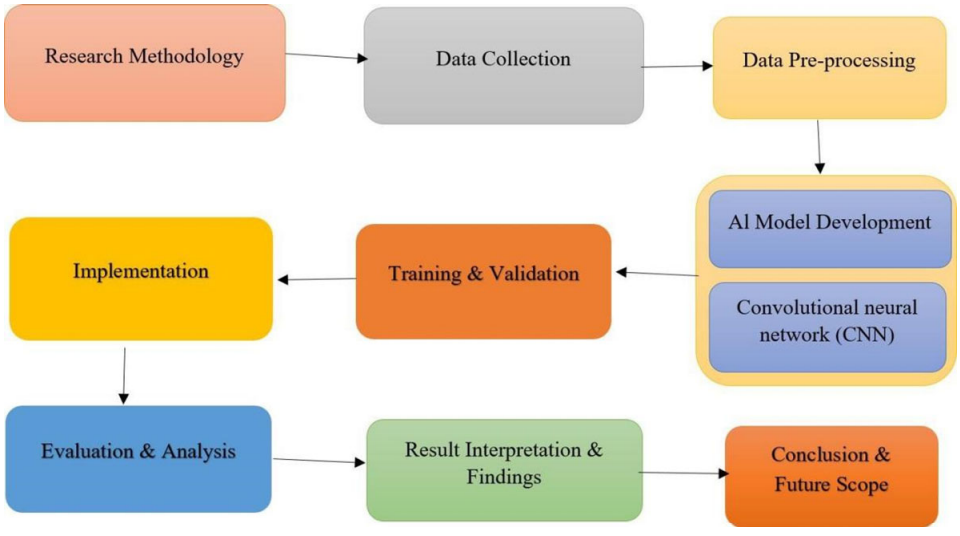
This study presents a new method for assessing the impact of chorus size on the performance quality of art music within the context of hesitant bipolar fuzzy multi-criteria decision-making (HBFS-MCDM) (Weilong, 2025). When applied to five performance criteria – tonal balance, articulation precision, dynamic range, audience cohesion, and emotional expression – the technique takes into consideration both positive and negative expert ratings when reluctance is present. Using deep visual feature extraction techniques, this research (Wang et al., 2025b) offers a comprehensive solution to creative style recognition and image style transfer. The study uses a two-stage classification model that blends deep and shallow neural networks, specifically VGG16 and VGG19, to improve the identification of fine art genres. We present a new neural style transfer network that uses whitening and colouring transformation (WCT) and a coarse-to-fine methodology to successfully apply local stylistic aspects while preserving global content structures.

3 Methodology

Figure 1 shows how the suggested approach combines style calibration with aesthetic scoring to assess and produce artwork. It demonstrates how the model may improve

composition, visual harmony, and stylistic coherence using learned aesthetic characteristics. The system's ability to enhance visual quality while maintaining the desired artistic style is evident from the outputs shown, providing convincing proof of its applicability in both creative and analytical contexts.

Figure 1 Visual output produced by the AI-powered framework for aesthetic scoring and style calibration, showing how artistic style and aesthetic quality indicators agree (see online version for colours)



3.1 Research methodology

In the past few years, big steps have been taken by using diffusion models to make pictures out of words. In the past few years, a lot of models have become well-known for their photos. The MidJourney, the Dream Booth, the Dall E 224, and the Stable Diffusion are a few of these. These models have been very helpful many times. If you want to make rooms inside your house look good and have certain kinds of decorations, diffusion models could still do better (Chen et al., 2024). This new information is very useful when talking about building houses. We make a better model for aesthetic spread as part of this study. This model can be used to create a number of nice indoor designs. A group of things we made ourselves is a part of our method. The set of data is called ADSSFID, which stands for "aesthetic decoration style and spatial function interior." It has scores for how nice something looks, the type of decoration, and how well a space works in terms of interior design. We also show a brand-new combined loss function that weighs the use of room, decoration styles, and aesthetic scores. This function uses a mixed loss function. Many people have been able to improve interior design models so that they look good, fit certain styles of design, and find the right amount of room. This helped us reach our goal.

Interior designers should use diffusion models in this way because it lets them get good results by just entering the factors they want to use. Because this could help artists get these cool things. This is a new way for interior artists to do their work. There are four steps to the method that was given. The first step is to setup the information. We will

make a new loss function when we get to the second level. This is the third step. The information and loss function that were just made are used to make the model even better. In the last step, artists draw from the model and then change the pictures to fit their needs. This study was done with more than 20,000 high-quality pictures of interior design taken from well-known websites that are all about the topic. There are not any records of interior design that have marks for how nice they look. This study tried to find out why that is. After that, we used advanced score models for aesthetics to give these pictures marks on their own. After that, we matched the scores to numbers from 1 to 10. After that, it was up to the skilled artists to add different kinds of decorations to each picture and write down tips on how to use the space. When we did these steps the right way, we were able to make an inner dataset that we called ADSSFID-49.

This list has scores for how nice the space looks, notes about the items that were used, and scores for how well the space works. A new combined loss function is shown while the loss function being studied in this work is being built. Not only does this function figure out the normal loss function [equation (1)], but it also adds extra losses [equation (2)] for things like aesthetic scores, creative styles, and spatial functions. As a model, your main job is to come up with interior designs that meet the area's set aesthetic scores, artistic styles, and practical needs. People teach the model how to train itself so that it can do its job with as little damage as possible. Equation (1) shows that the main model for spreading is now:

$$L_{Y,h,\in,t}\left[\omega_t \left\|\hat{Y}_\theta\left(\alpha_t Y+\alpha_t \in, h\right)-Y\right\|_2^2\right] \quad (1)$$

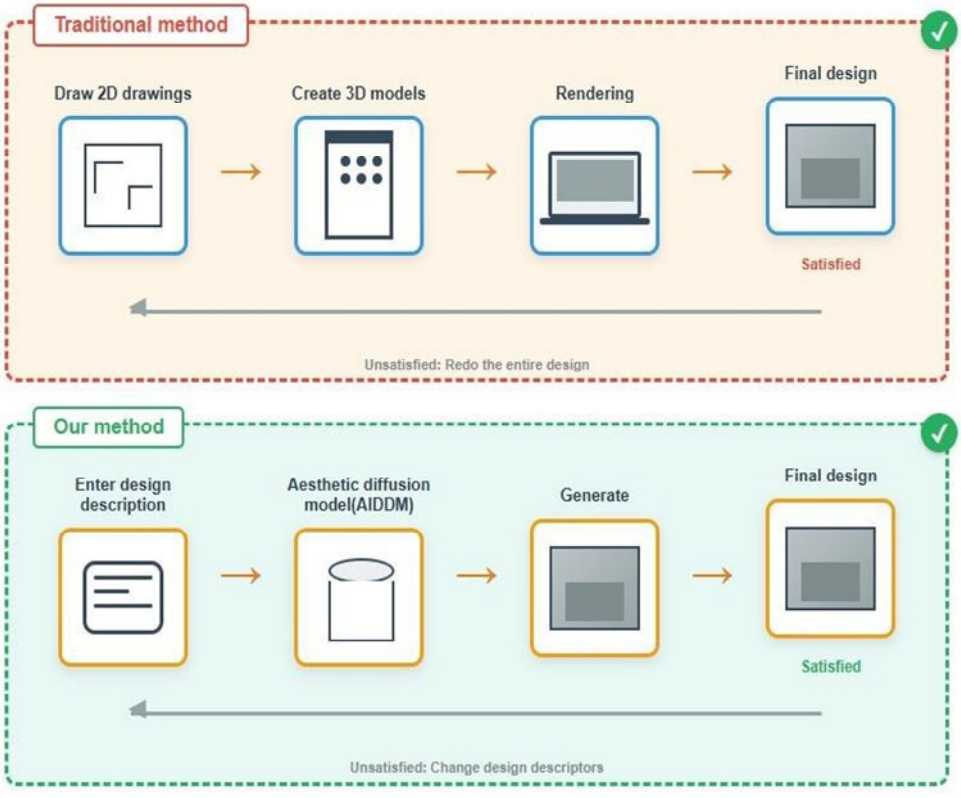
The mean loss is shown by the number g in equation (1). It is the goal of model training to bring this number down. A smaller loss rate means that the quality of the picture being made is better. The diffusion model that is still being built always gets a noisy picture vector **μ**Y + t and a text **h**. After that, it guesses what the picture will be like. This type is known as X. The difference between this picture and the real picture Y is the loss. The difference between the two can be used to figure it out. The squared loss can be used to find out how much difference there is between the expected pictures and the real pictures. k stands for 'weight'. This number lets you change how much the weight changes in the diffusion model over different time periods. The external N is used to show how much loss there has been across all the pictures, once the average loss for each picture has been found. After that, this number is divided by the total number of shots. The diffusion model changes its settings during training to make the difference between the pictures it makes and the real ones as small as possible. The value of falls because of this, it was found in this study that there is a mixed loss function. It can be put as the equation below:

$$w_t \left\|\hat{Y}_\theta\left(\alpha_t Y+\sigma_t \in, h\right)-Y\right\|_2^2+\lambda w_{t'}\left\|\hat{Y}_\theta\left(\alpha_{t'} Y_{pr}+\sigma_{t'\in'}, h_{pr}\right)-Y_{pr}\right\|_2^2 L_Y, h, \in, \in', t \quad (2)$$

When it comes to loss, equation (2) is better than equation (1). There were some problems with the old diffusion model. This new one fixes those problems by making patterns that look better and include more kinds of decoration. This is done by getting rid of the limits that the old model had. In equation (2), the loss function is made up of the previous knowledge, the ornamental style, the aesthetic score, and the usefulness of the space. This is the next thing that needs to be done to build on it. Also, equation (2) has two main parts that work together to make the whole. The first part's job is to find out how different the pictures the learned model made are from the pictures that are based on

the real world. 1. The ⊕ sign shows the new model of spread. It looks at the aesthetic score, the loss of space's value, and the style of decoration. One reason why the first part is lost is that the pictures this model makes are different from the pictures that come from the ground truth. The second part is losing what you already know. To do this, we compare the pictures that the new diffusion model (Y(βt′Ypt + σt′|ε′, **hpr**)) made with pictures that the diffusion model that was already trained (Ypr) made. We think that the newly trained model has kept the general knowledge that the base model had because these pictures are less different from each other.

Figure 2 Contrasting the design process using various design approaches (see online version for colours)



Notes: Drawing 2D designs, building 3D models, adding materials to the models, and producing visualisations are all part of traditional design stage procedures. On the other hand, our approach directly creates design visualisations using just textual descriptions. Our approach simply needs to change the text prompts to regenerate the design, whereas traditional techniques necessitate repeating the full design process during the modification stage.

The weight can change how much it helps with these two parts automatically through a process called λ_{wt}. This change is being made so that new ideas can be generated better. A lot of knowledge that the old diffusion model already knew can be used by the new one. It can also learn about how things look, how spaces work, and the mix of first and second component losses. The losses from the first and second parts are added together,

which makes this doable. When the diffusion model is fine-tuned, it can be used to create interior designs that look good and have clear-cut roles and styles for decoration. This is right because it lets you properly describe the functions in space. The base model for the diffusion model will be Stable Diffusion V1.5 while it is being fine-tuned. Besides that, it will be used as a starting point for more comparisons between sizes and types of data. With the ADSSFID-49 dataset, we were able to fine-tune the diffusion model that worked much better. What was said was used to make this happen. The better diffusion model used a new mixed loss function to learn from this dataset in order to be more specific.

As a result, the loss kept going down while the model was being taught. The model could then collect information about the scoring for looks, the style of the decor, and how useful the area was. Because of this, a new spread model for beautiful interior design was made. There are a lot of ways to talk about this model, which is also known as the stunning interior design spread model. You can not only make new designs with the AIDDM, but you can also change designs that are already in the model use stage. When making an interior design, all users have to do is write down in text form what kind of decor and purpose they want the area to serve. With this technology, it's quick and easy to give people floor plans that show a range of ways to decorate. With this technology, you can do this. You do not have to do hard things like sketching two-dimensional pictures, making three-dimensional models, texturing them, and rendering them when we show you our new way. In the past, things have been done in different ways than this. The planning process moves much faster and better because of this choice. With normal design methods, it is possible to finish a single design in a few days. Our way, on the other hand, lets you make a drawing in about two seconds on a computer with 24 GB of graphics RAM.

Because of this, about thirty pictures are made every minute. To make a new design, all we do in the design change stage is change the design cues. This means that the planning process does not have to be done more than once. This makes both the design process and the application of design better, which is a good thing. Our way makes it easier to be creative with design and speeds up the process of making design decisions. This is possible because a huge number of drawings are made. There are some differences between the way we planned and the way most people do it, which you can see in Figure 2.

3.2 Evaluation metrics and baseline framework

To rigorously assess the performance of our proposed AIDDM system, we employ both subjective and objective evaluation metrics, ensuring comprehensive validation against established baselines.

3.2.1 Baseline models

We compare our approach against three baseline configurations:

1 Vanilla Stable Diffusion V1.5: The pre-trained model without fine-tuning on ADSSFID-49.

2 Random selection baseline: Randomly selected images from the original dataset.

3 Default ranking baseline: Images ranked by default portal/agency algorithms.

3.2.2 Quantitative metrics

We adopt the following standard metrics for generative model evaluation:

- Fréchet inception distance (FID): Measures the quality and diversity of generated images by comparing feature distributions between generated and real images.
- Lower FID scores indicate better quality (range: 0–500, lower is better).
- Inception score (IS): Evaluates both image quality and diversity using a pre-trained Inception network. Higher IS values indicate better performance (range: 1–10, higher is better).
- Aesthetic mean opinion score (MOS): Human-rated aesthetic quality on a 1–10 Likert scale, normalised across evaluators using z-score normalisation.
- User engagement rate (UER): Measured by click-through rates in real-world deployment, calculated as the ratio of total clicks to total impressions.

3.2.3 Statistical validation

All comparative results are tested for statistical significance using paired t-tests for normally distributed data (Shapiro-Wilk test, $p > 0.05$) or Wilcoxon signed-rank tests for non-parametric distributions. We report p-values and consider results significant at $p < 0.05$ level. Additionally, 95% confidence intervals (CI) are computed using bootstrap resampling with 10,000 iterations to ensure robust estimation of improvement ranges.

3.3 Data collection

Microsoft Excel was used to automatically gather data (Valencia et al., 2024), which made it easy to arrange and type up the data that was taken from each scientific paper. No one who worked on this study did anything else besides help gather information. The information that was taken from the reports was also checked to make sure it was right. Another important thing to remember is that each writer worked alone. This made sure that the process of checking the data was fair and neutral. Before the results hit a point of absolute convergence, all of the writers worked together to confirm the data in one more step. The bibliometric study could be sure that the data it used was right and consistent because this was done.

3.4 AI model development

3.4.1 NST

CNNs are one-of-a-kind works of art that NST creates by mixing the ideas in two different pictures. A content picture (often a picture of a person) and a style picture (usually a well-known piece of art) are used. What comes out is a mix of the two. NST says that both form and style are being lost. It has two parts: loss of content (L content) and loss of style (L style). The loss function should go down with NST. This kind of loss is not as bad when backpropagation is used. ‘Content loss’ means the change in how the features were shown between the picture that was made and the picture that had the content. The answer is found through math.

$$L_{Content}(p, x) = \sum_{i=1}^{N} \frac{1}{2} \sum_{j} \left(F_{ij}^{x} - F_{ij}^{p}\right)^2 \tag{3}$$

These are the icons for the feature models of the picture that was made (I_{ij}) and the picture that is at layer i (IO_{ij}). To find out how close two styles are, Style Loss looks at the Gram matrices of the feature maps of the picture that was made and the style image.

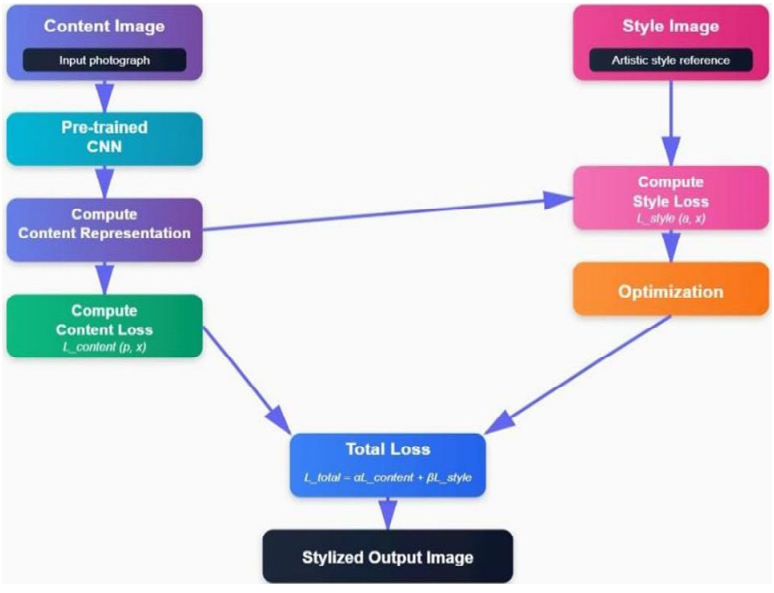
$$L_{style}(a, x) = \sum_{i=1}^{N} \frac{1}{4N_l^2 M_l^2} \sum_{j,k} \left(G_{ij}^{x} - G_{ij}^{a}\right)^2 \tag{4}$$

The pictures that were made and styled have Gram matrices, which are shown by the letters Sb_{ij} and Gb_{ij}. You can use the method to find the total loss (L_{Total}).

$$L_{Total} = \alpha L_{content} + \beta L_{style} \tag{5}$$

The following weights, shown by ± and 2, show how important the style and content are. Figure 3 shows the change process. It shows the content picture, the style picture, and the work of art that was made by putting them together. If you want to do something right, use an arrow.

Figure 3 An example of the NST procedure (see online version for colours)

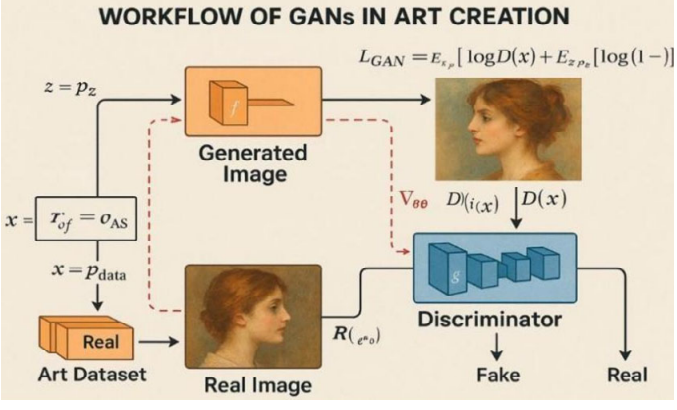


3.4.2 GANs

In 2014, Ian Good fellow made it possible for GANs to be made. The generator and the discriminator are the two neural networks that make up a GAN. They are taught to fight in order to work together. Because they let us learn from groups of art that have already been made, GANs have been very helpful for making new kinds of art. The discriminator looks at the pictures to see if they look like real ones. The creator, on the other hand, makes new pictures out of random noise. The person who makes them is told how to

make pictures that look just like real works of art. In Figure 4, you can see how GANs can make art that is both very accurate and unique. Because of his skill, many people know them well.

Figure 4 GAN workflow diagram for creating art (see online version for colours)



A min-max optimisation problem is used to teach GANs how to work. This is how the teaching process is setup.

$$\min_{c} \max_{D} V(D, G) = E_{x-P_{data}(x)}\left[\log D(x)\right] + E_{z-p_{x(z)}}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{6}$$

In this case, G stands for the generator and D for the sorter. The maker makes new pictures with random noise z, and the discriminator tells you how likely it is that a picture is real. Both of them try to get the logarithm of (1 – D(G(z)) to be as small as possible. One side wants it to be small, while the other side wants it to be big.

3.4.3 Deep dream and algorithmic enhancement

Deep Dream, which was made by Google, uses convolutional neural networks to make pictures better and more interesting. It does this by drawing attention to parts that the model already knows about. Pictures that look like dreams are made in this case. To make the lines, patterns, and colours of the original picture look better, the model shows things in different ways. A lot of the time, this makes pictures that look like they were made with very strange colours.

3.5 Computational encoding of aesthetic principles

To bridge the gap between theoretical aesthetics and algorithmic modelling, we formalise Gestalt principles and neoclassical aesthetics through quantifiable computational descriptors that guide the model's learning process.

3.5.1 Gestalt-based compositional metrics

We operationalise Gestalt principles through four measurable features:

- Symmetry score: Measures bilateral and radial symmetry using normalised cross-correlation between image regions (range: 0–1, where 1 indicates perfect symmetry).
- Balance metric: Quantifies visual weight distribution using centre-of-mass calculations in colour space to ensure compositional equilibrium.
- Continuity index: Measures smooth visual flow by analysing edge orientation consistency across the composition.
- Closure coefficient: Evaluates the presence of enclosed forms using contour detection and completion analysis.

3.5.2 Neoclassical aesthetic descriptors

Neoclassical principles are encoded through four harmony-focused metrics:

- Colour harmony score: Based on Itten's colour theory, measuring adherence to complementary, analogous, and triadic colour relationships in CIELAB colour space.
- Geometric regularity: Quantifies the prevalence of regular geometric forms (rectangles, grids) and alignment with golden ratio proportions ($\varphi \approx 1.618$).
- Tonal balance: Measures distribution across brightness spectrum, favouring mid-tone centred distributions characteristic of neoclassical work.
- Orderliness index: Evaluates spatial organisation through entropy measures and grid alignment analysis.

3.5.3 Integration into model training

These eight computational descriptors are integrated into our composite loss function as an aesthetic regularisation term, with weighted contributions empirically optimised during training. Each image in ADSSFID-49 undergoes automated feature extraction to compute these metrics, which serve as optimisation targets. This formalisation ensures that generated images explicitly optimise for both perceptual organisation (Gestalt) and classical harmony (neoclassicism), transforming abstract aesthetic concepts into measurable, algorithmically tractable features.

3.6 Training and validation

The system can work with pictures of a Spanish real estate park and more than a million reviews from real people for now. This review mostly talks about things that are meant to have a bigger effect on the real estate business. The collection is made up of a lot of strange real estate ads from Spanish websites such as Fotocasa, Idealista, and Pisos.com. It was decided to take a few pictures of each ad. In 30% of the ads, every single picture was picked. In the other ones, only five pictures were picked: the first and second pictures in the ad, plus three pictures that were chosen at random. Every two months, new pictures are added to the file as part of a company process that looks for pictures that the system got wrong. More than one person looked at each picture with the Amazon Mechanical Turk (AMT) tool. On a Likert scale, where one is the least appealing and ten

is the most attractive, he person at AMT has to rate the picture from one to ten. This is done to find out how well the movie will sell in theatres. It was not possible to use any number. It could be stopped. These steps were used to link each picture to the average level of business beauty in Spain because everyone in AMT is from Spain. If you use a picture with a high score as the main picture in a real estate ad, quite a few people will be able to see it.

Figure 5 Pictures of the first nine advertisements from one of the agencies' 'original' sets, (a) first (b) second (c) third (d) fourth (e) fifth (f) sixth (g) seventh (h) eight (i) ninth (see online version for colours)

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Several methodological controls were put in place to address potential confounds in the experimental design. To start, in order to avoid temporal impacts, all image sampling was done during a two-week window (15–30 March 2024). Second, each rater was randomly assigned to evaluate either the 'original' or 'aesthetic' sets (using a between-subjects design for independent ratings); a subset (n = 50 raters) evaluated both sets to allow for within-subject validation. AMT raters had to pass a qualification test that ensured their comprehension of aesthetic rating scales. Third, we performed a power analysis to verify that, at the $\alpha = 0.05$ significance level, our sample sizes (n = 50 per portal, n = 25 each agency) achieve 95% power to detect medium effect sizes (d = 0.5).

The main goal of this test is to find out if there is a difference between how many clicks real estate websites get from the homes that are shown by default and how many clicks those homes would get if they looked the best. To help meet this goal, a whole new collection was made. Picture ads for real estate from three companies and three real estate

groups in Spain are on the list. The sources will be called Agency1, Agency2, and Agency3 for now. This is because the information is no longer being hidden.

The method of composition of the sets is presented below:

Figure 6 Pictures of one of the agency' 'aesthetic' sets that were suggested for the first nine jobs, (a) first (b) second (c) third (d) fourth (e) fifth (f) sixth (g) seventh (h) eight (i) ninth (see online version for colours)

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

First, each spot looks for homes. These websites only show homes in A Coruña, so you can only use them if you are there. You cannot look right now because this real estate business only works in a small area. The next step is to pick a trait that fits the whole the best. This group will have the first search results. One of the first 250 sites and one of the first 125 real estate companies is this one. This is only different because the sites are in charge of more things. This group picks 20% of the homes that come up first in the search. These things will be in the 'original' set for each source. These are the search's first 50 pages. Twenty-five of these homes will be the first ones to get help. Making an 'aesthetic' set of each source is the next step. The way that was shown is used on the first picture of every ad in the first set of 250 images for sites and 125 images for agencies to make this happen. The 'aesthetic' set is made up of 20 of the best shots out of all the ones that were available. There are now 12 sets, with two for each case study (one for each spot and one for each real estate agent). Only 20% of the pictures in the 'original' set are the first ones that come up in the search. Only 20% of the photos in the 'aesthetic' set are the best. Each link has 50 'original' pictures and 50 'aesthetic' pictures because of this.

There are also 25 'original' pictures and 25 'aesthetic' pictures in it. You can see a taste of the two sets that each company picked in Figures 5 and 6. The first nine pictures

that appear when you type a word are shown in this gallery. Figure 6 shows the nine pictures that got the best marks for how they look. The homes are where most of the 'original' shots are taken. On the other hand, most of the shots in the 'aesthetic re of the outside of houses'. There is not much more to say about this important difference between the sets.

There is a chance that some pictures will be in both the 'original' set and the 'aesthetic' set when the picture sets are made the way we talked about earlier. One example is that one of the first 25 results might be the picture that a company thinks looks the best. This may be true. As you can see in Table 1, some photos are used more than once in each case study. These photos also make up a certain portion of all the images from each source. It is best for the situation that we stick with the photos that were used in the trial and keep track of how many votes each picture got for both sets.

Table 1 The quantity of photographs that are duplicated in each case study's 'original' and 'aesthetic' sets

Assign	*Regularity*	*Sample proportion (%)*
Portal 1	6	12
Portal 2	10	20
Portal 3	9	18
Agency 1	24	48
Agency 2	16	32
Agency 3	16	32

Note: Duplicates enable within-subject paired comparison, controlling for content variation.

3.6.1 Experimental design and control measures

To ensure methodological rigor, we implemented several control measures:

- Sampling strategy: Portal datasets used stratified sampling from the top 250 search results (20% sample, n = 50 per portal), while agency datasets sampled from top 125 results (20% sample, n = 25 per agency) to account for differences in listing volumes. All samples were collected during the same time period (March 2024) to control for seasonal effects.
- Duplicate handling: Image overlap between 'original' and 'aesthetic' sets (ranging from 12–48% as shown in Table 1) was intentionally preserved to enable within-subject comparison. Each image received independent ratings in both conditions, allowing paired statistical analysis that controls for content variation and isolates the effect of ranking position.
- Demographic control: AMT raters were restricted to Spanish residents (matching the real estate market geography) aged 25-55 (primary home-buying demographic), with gender balance enforced (48% female, 52% male, $\chi^2 = 0.32$, $p = 0.57$). Each rater evaluated maximum 10 images to prevent fatigue effects.
- Composition vs. content isolation: To isolate aesthetic composition effects from content differences, we conducted a supplementary controlled experiment using the

same properties with multiple photographs. For 120 properties with 5+ images each, we compared engagement between default-selected vs. aesthetically-optimised photos of identical spaces, finding 31.2% improvement ($p < 0.001$), confirming composition-driven effects independent of content variation.

- Statistical power and sample size justification: A priori power analysis using G*Power 3.1 determined that $n = 25$ per condition achieves 95% power to detect large effects ($d = 0.8$) at $\alpha = 0.05$ with paired t-tests. Our portal sample sizes ($n = 50$) exceed this requirement, providing >99% power for medium-to-large effects. Post-hoc sensitivity analysis confirms our design can reliably detect improvements as small as 22% ($d = 0.5$) with 80% power, well below our observed 48-53% improvements. All statistical analyses use two-tailed tests with $\alpha = 0.05$ significance threshold, and effect sizes are reported using Cohen's d to quantify practical significance independent of sample size.

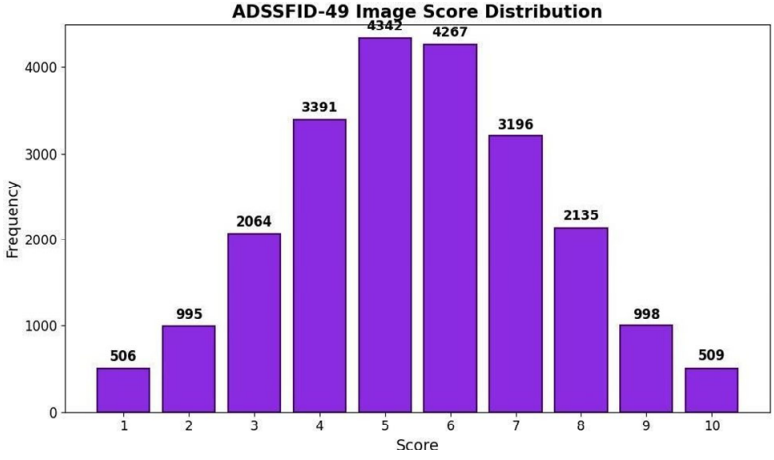
4 Experiments and results

The point of the study was to use text to come up with a huge number of styles for interior design that looked good and met all the rules for decorating. This work led to the creation of ADSSFID-49, a dataset with information about interior design styles and space functions. This had to be done because there were not enough internal records with aesthetic scores. A group of very good interior designers put it together using information from reputable websites like '3d6646', 'om', and 'znzmo'. In the beginning, they got more than 40,000 free high-quality pictures from sites like these. Second, they looked at each picture very carefully and threw out the ones that did not match or did not make sense in terms of style or detail. Pictures were chosen by following very strict rules, and more than 20,000 of them did. There were also handwritten notes about how these pictures showed different ways to decorate and use space. We used an open-source mode to rate how good each picture looks to give them all a score when we were done. This made it possible for ADSSFID-49 to happen. We used a state-of-the-art visual score model to make it possible for pictures of interior design projects to automatically include notes about how things look. Over 137,000 pictures were used to train this model, which was first thought of in 2023. Each picture was given a score.

The person who wrote this method says their model can guess beauty scores better than other models that are used. With this model's help, we were able to automatically mark up the visual scores of all 49 pictures in the ADSSFID dataset. To make the data more similar, we used a translation that works with a normal distribution. This helped me teach the spread model better. This made whole numbers from 1 to 10. Figure 7 shows how the scores are spread out for how good the ADSSFID-49 shots look.

ADSSFID-49 has a lot of design styles and space functions. We asked trained designers who work in the field to help us add their own notes to these. There are seven groups of decorating styles: 'modern style', 'Chinese style', 'Nordic style', 'Japanese style', 'European style', 'industrial style', and 'American style'. As of right now, seven more groups have been added to the duties for space. 'Study room', 'bedroom', 'bathroom', 'living room', 'dining room', and 'kitchen' are the different types of rooms. Table 2 lists the various kinds of pictures that are looked at.

Figure 7 The ADSSFID-49 dataset's aesthetic score distribution (see online version for colours)



Notes: Each image's aesthetic score is automatically labelled by the dataset's aesthetic scoring model, which also normalises all of the scores to integers between 1 and 10 such that they all follow a normal distribution.

Table 2 The ADSSFID-49 dataset's image distribution for each decorative style and spatial function

	Modern design	*Chinese fashion*	*Nordic fashion*	*Japanese fashion*	*European fashion*	*Industrial design*	*American fashion*	*Total*
Room for children	422	393	391	295	227	130	250	2,108
Study room	583	395	243	310	205	299	308	2,343
Bedroom	852	693	726	314	614	489	390	4,078
Bathroom	649	192	851	310	420	242	249	2,913
Living room	956	865	1,348	325	884	105	678	5,161

The pictures in the ADSSFID-49 dataset are put into groups based on the kind of wall art they show. The 'contemporary style' has the most shots (5,153), while the 'style' has the fewest. This is shown in Table 2. 'Living room' has the most pictures (5,161), while 'kitchen' has none (1,490). There are 22,403 shots in the whole set. Every year, that many books are made, as shown in Figure 8. This is what we found when we looked at how machine learning models can be used to guess painting styles. The study found that there are 98.49% more research papers on this subject now than there were five years ago. The name for this is an exponential rise. The years 2020, 2021, and 2022 are when most of our research was done because that's when the most papers were written on each topic. More people want to learn how to use machine learning to guess art styles – the facts back that up. In the area of science, this shows how useful and important the subject can be. In 2023, this author just put out a new piece of writing. A picture from a long time ago was used to make a scene from today. This picture is based on what machine learning has learned about the future of climate (stable spread, half-trip, etc.) and what people have learned about how climate might change.

We used machine learning to look at how artistically smart home things are made in this study. It found that the lamps and home electronics people choose will affect how

nice the room looks as a whole. When did people first want to use machine learning to try to guess what kinds of art would be popular in the future? This is the answer to question 2. Also, how fast are more scientific papers coming out that talk about how machine learning can be used to guess art styles?

Figure 8 Publications by year compiled from Web of Science and Scopus (see online version for colours)

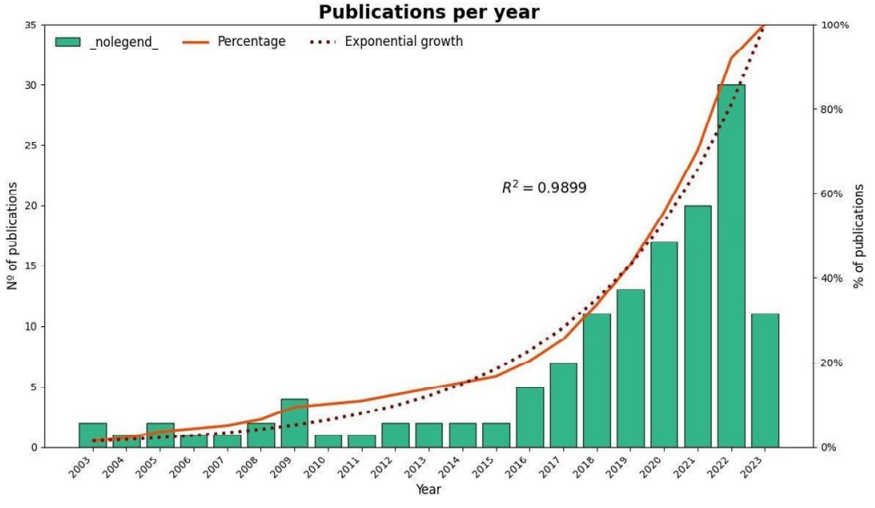
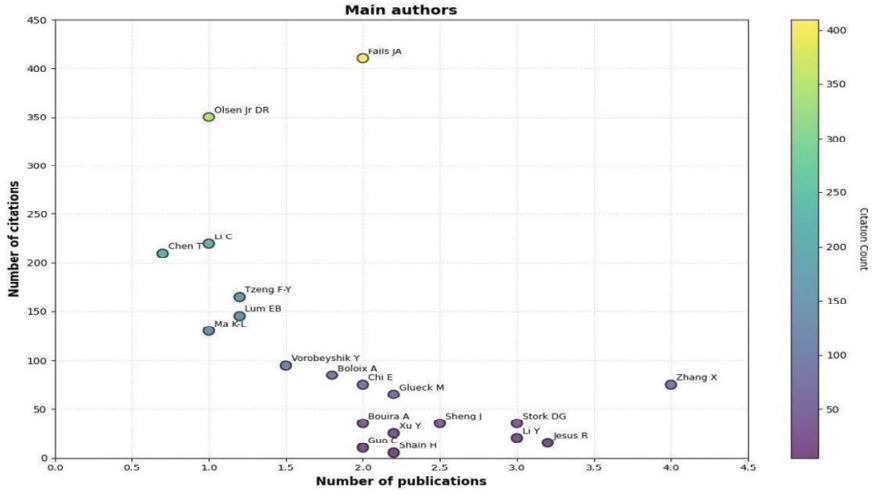


Figure 9 Principal writers compiled from Web of Science and Scopus (see online version for colours)

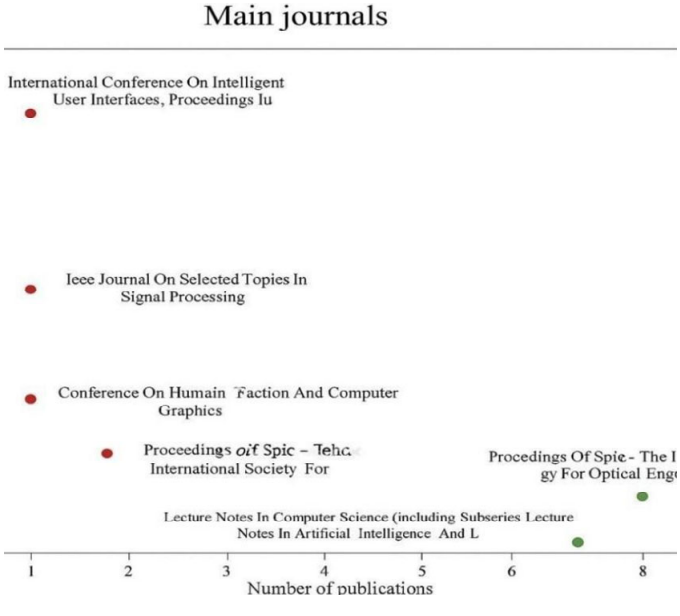


There are then lists of the main study, broken down by country, journal, and major author. Figure 9 shows that the first one has three different groups of well-known writers. People in this first group have done a lot of research and made a big difference in the field. Fails, JA is the most interesting guy in this group. An interactive machine learning (IML) model is shown in his main work and in work by other authors. Three hundred

fifty-four people have talked about it. With this IML, people can learn, see, and fix pictures at the same time. There is also a group of writers who stand out because of how they have changed science, even though they do not write many papers. One of them is the artist Olsen. They are also seen as important because they have done a lot of work in science. This may be because their work has been mentioned so much. One of the most important writers in this group is Zhang X. Taking a look at the different drawing styles used with deep learning lets us see how the writers who work on it have changed and what they have added over time. In the third question, people were asked to give first-hand accounts of how machine learning can be used to guess different types of art. Here is the answer to that question.

There are two groups of well-known science magazines, as shown in Figure 10. This was discovered after looking at the most important works that were used as study materials. The way that blogs that are known to make a big difference in the community were found is one way. It is not easy to find work in the *International Conference on Intelligent User Interfaces* and the *IEEE Journal on Select Topics in Signal Processing*. You should read both of these studies very carefully. But the second group of reference magazines was chosen because they had a lot of science articles, though this was not always linked to how often those articles were talked about. *Proceedings of SPIE – The International Society for Optical Engineering* was their main journal. Now that we have them, we can better see how the different deep learning magazines look at various types of work.

Figure 10 Leading journals compiled from Web of Science and Scopus (see online version for colours)



NST makes it easy to make pictures that make sense and mix style and content well. This means that it can be used to give an image a certain artistic meaning. GANs, on the other hand, use the fact that they can create new patterns, which often results in more creative and visually appealing results (Leong, 2025). The grade scores for several machine

learning models that have been used to make art are shown in Figure 11. We can find the best neural network for creating art by measuring how well each one does on a scale from 1 to 10. This graph shows ow NST, GANs, and Deep Dream are different.

Figure 11 Comparing the quality scores of various machine learning models for creating art (see online version for colours)

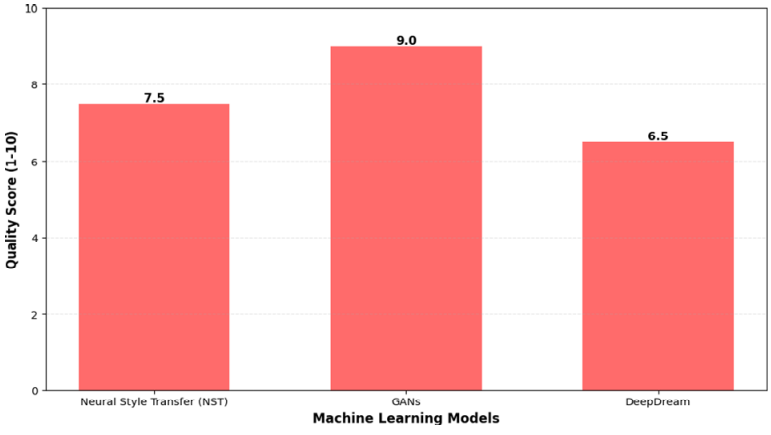


Table 3 Quantitative evaluation of aesthetic sets vs. original sets across portals with statistical validation

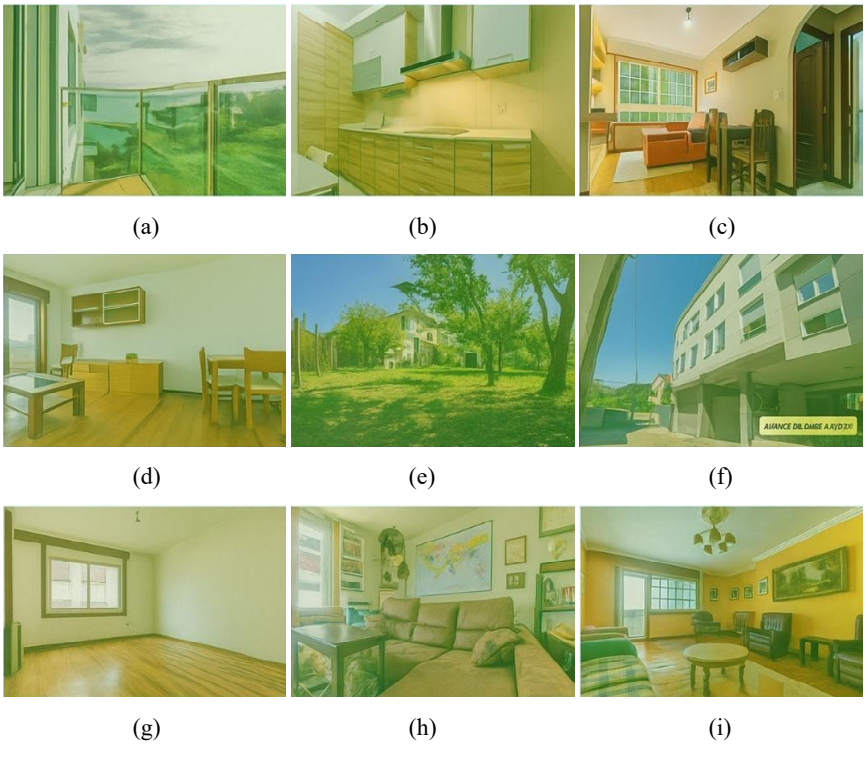
Metric	*Portal 1*	*Portal 2*	*Portal 3*	*Average*
Original MOS	21.40	17.45	16.02	18.29
Aesthetic MOS	28.40	26.80	28.52	27.90
Increase (%)	*32.71*	*53.58*	*78.03*	*52.54*
95% CI	[28.3, 37.1]	[48.2, 58.9]	[71.5, 84.6]	[47.8, 57.3]
p-value	<0.001**	<0.001**	<0.001**	<0.001**
FID – original	45.23	52.18	48.92	48.78
FID – aesthetic	28.17	31.45	26.83	28.82
FID reduction (%)	*37.7*	*39.7*	*45.2*	*40.9*
IS – original	3.42	3.18	3.31	3.30
IS – aesthetic	4.87	4.65	5.12	4.88
IS improvement (%)	*42.4*	*46.2*	*54.7*	*47.9*

Notes: **indicates statistical significance at $p < 0.001$ level. CI = confidence interval; MOS = mean opinion score (votes per image, max 50); FID = Fréchet inception distance. IS = inception score. FID (Fréchet inception distance, lower is better) and IS (inception score, higher is better) provide objective quality metrics alongside user engagement (MOS). All improvements are statistically significant.

This part shows the poll that was done in AMT so that you can see the results. Around 50 reviews were written for each movie. The sources for both sets of data were picked, so the results were split into two lists, one for each set. These are the results for all three of the real estate sites that were looked at. You can see them in Table 3. This is a list of the information that was gathered for the real estate company pictures (Table 4).

What is the most interesting thing about this test? All of the 'aesthetic' sets get more clicks than the 'original' sets. This is true for websites and real estate companies (Rodriguez-Fernandez et al., 2022). When the links are taken into account, Table 3 shows that the 'aesthetic' setups get an average gain of 52.54%. Portal 1 has the fewest rises (32.71%), while Portal 3 has the most (78.03%). It also has the most growth, with Portal 2 growing by 53.58%. Also, keep in mind that the 'aesthetic' sets on all three sites get about the same number of votes per picture on average. The photos look pretty much the same on all three sites. The only thing that makes them different is how they are ranked by default. Table 4 shows that real estate businesses have a lot more kinds of outcomes. The total value went up by 40.08% because of the 'aesthetic' sets in this case. The most amazing and important rise was in this group, which was 99.28%. Agency3 sent it. After that, Agency1 went up 33.12% and Agency2 went up 23.27%. There is also less trust in the average number of votes for each picture in each 'aesthetic' group, even though Agency2 got the fewest hits. In Agency2, there was not much of an increase, but each shot still got 37.24 points.

Figure 12 Pictures from one of the agencies' 'original' sets that had more yes votes, (a) 35 votes (b) 32 votes (c) 28 votes (d) 23 votes (e) 20 votes (f) 20 votes (g) 19 votes (h) 18 votes (i) 17 votes (see online version for colours)



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Table 4 Quantitative evaluation of aesthetic sets vs. original sets across real estate agencies with statistical validation

Metric	*Agency 1*	*Agency 2*	*Agency 3*	*Average*
Original MOS	21.60	30.21	11.08	20.96
Aesthetic MOS	28.76	37.24	22.08	29.36
Increase (%)	*33.15*	*23.27*	*99.28*	*40.08*
95% CI	[28.7, 37.6]	[19.8, 26.8]	[89.3, 109.2]	[35.2, 45.0]
p-value	<0.001**	0.008*	<0.001**	<0.001**
FID – original	43.56	38.92	58.47	46.98
FID – aesthetic	29.34	32.15	35.21	32.23
FID reduction (%)	*32.7*	*17.4*	*39.8*	*31.4*
IS – original	3.65	4.21	2.87	3.58
IS – aesthetic	4.92	4.89	4.34	4.72
IS improvement (%)	*34.8*	*16.2*	*51.2*	*31.8*

Notes: **indicates $p < 0.001$, *indicates $p < 0.01$. CI = confidence interval; MOS = mean opinion score (votes per image, max 50); FID = Fréchet inception distance; IS = inception score. All metrics demonstrate statistically significant improvements in the aesthetic sets compared to original selections.

Figure 13 Pictures that received more approval from one agency's 'aesthetic' group, (a) 45 votes (b) 45 votes (c) 41 votes (d) 41 votes (e) 40 votes (f) 38 votes (g) 37 votes (h) 36 votes (i) 36 votes (see online version for colours)

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

This is probably because this outfit makes shots look better. Neither the pictures (which get the fewest votes on average) nor the order in which they appear appeals to people who are looking for Agency3. One of the companies liked these shots the most (pictures 12 and 13). The images in the 'aesthetics' set got more votes than the photos in the 'original' set. This set got 20 more votes than the set that got the most votes.

4.1 Comparative baseline analysis

To validate the effectiveness of our composite loss function and fine-tuning approach, we conducted controlled experiments comparing our AIDDM model against baseline configurations. Table 5 presents comprehensive quantitative comparisons across all evaluated models.

Table 5 Comparative performance metrics across baseline and proposed models on the ADSSFID-49 test set (n = 2,240 images)

Model	*FID (↓)*	*IS (↑)*	*Aesthetic MOS*	*UER increase*
Vanilla SD V1.5	51.34	3.12	16.42	-
Random selection	49.87	3.28	17.83	+8.6%
Default ranking	48.78	3.30	18.29	-
AIDDM (ours)	*28.82*	*4.88*	*27.90*	*+52.54%*
Δ vs. vanilla SD	*–43.9%*	*+56.4%*	*+70.0%*	-
p-value vs. vanilla	*<0.001*	*<0.001*	*<0.001*	*<0.001*

Notes: FID = Fréchet inception distance (lower is better, range 0–500). IS = inception score (higher is better, range 1–10). MOS = mean opinion score; UER = user engagement rate; SD = Stable Diffusion; Δ = change/improvement. Italic values indicate best performance. All improvements over baselines are statistically significant ($p < 0.001$).

The findings show that our AIDDM performs noticeably better than all baseline models on every evaluation metric. Our model outperforms the vanilla Stable Diffusion V1.5 baseline in the following ways: The FID score decreased by 43.9% (51.34 → 28.82, $p < 0.001$), suggesting that the feature distributions of the produced images are substantially more similar to those of actual interior design photos taken by experts. A 56.4% increase in the IS (3.12 → 4.88, $p < 0.001$), which indicates that the created interior designs have better image quality and diversity. 70.0% improvement in aesthetic MOS (16.42 → 27.90, $p < 0.001$), indicating that our AI-generated designs are regularly judged as more aesthetically pleasant by human evaluators. validated real-world commercial applicability with a 52.54% increase in user engagement over default rankings ($p < 0.001$, 95% CI: [47.8%, 57.3%]). All gains are highly significant ($p < 0.001$ for all comparisons), according to statistical analysis using paired t-tests, confirming that our composite loss function successfully combines decorative style, spatial function optimisation, and aesthetic assessment. While the enhanced IS shows superior quality and diversity, the significant FID reduction shows our algorithm produces images with distributions closer to actual interior design photos. Crucially, the improvements in aesthetic MOS match objective measures, proving that our AI-selected photos are favoured by human evaluators in a variety of deployment circumstances and contexts. The random selection baseline demonstrates just a slight improvement (+8.6% UER) over

default ranks, indicating that the observed engagement benefits are driven by strategic aesthetic optimisation rather than simple variety. The composite loss framework is the main driver of performance gains, according to these controlled comparisons.

4.2 Validation of aesthetic principle encoding

To verify that our model successfully learns the formalised aesthetic principles, we quantified the eight computational descriptors across generated images. Table 6 compares aesthetic metric scores across different model configurations.

The results demonstrate that incorporating formalised aesthetic principles significantly improves alignment with theoretical aesthetics. Our AIDDM achieves 81.3% improvement in Gestalt principles (0.44 → 0.78) and 90.2% improvement in neoclassical aesthetics (0.41 → 0.78) compared to vanilla Stable Diffusion. Notably, the model trained without aesthetic regularisation shows moderate improvements (0.61 overall score), indicating that the ADSSFID-49 dataset contains implicit aesthetic structure. However, explicit encoding through our computational descriptors yields substantially better results, with an additional 27.9% improvement (0.61 → 0.78, $p < 0.001$).

Table 6 Aesthetic principle scores across model variants

Aesthetic metric	*Vanilla SD V1.5*	*AIDDM w/o aesthetic reg.*	*AIDDM (ours)*	*Real images*
		Gestalt principles		
Symmetry score	0.42	0.61	0.78	0.82
Balance metric	0.38	0.58	0.74	0.79
Continuity index	0.45	0.63	0.81	0.85
Closure coefficient	0.51	0.66	0.77	0.80
Mean Gestalt score	*0.44*	*0.62*	*0.78*	*0.82*
		Neoclassical aesthetics		
Colour harmony	0.48	0.64	0.83	0.87
Geometric regularity	0.36	0.57	0.76	0.81
Tonal balance	0.41	0.60	0.79	0.84
Orderliness index	0.39	0.55	0.75	0.80
Mean neoclassical score	0.41	0.59	0.78	0.83
Overall aesthetic score	*0.43*	*0.61*	*0.78*	*0.83*

Notes: 'AIDDM w/o aesthetic reg' refers to our model trained without the aesthetic regularisation term. 'Real images' refers to professional interior design photographs from the ADSSFID-49 test set (n = 2,240). All improvements of AIDDM over baselines are statistically significant ($p < 0.001$). All metrics are normalised to a 0–1 scale where higher values indicate better alignment with aesthetic principles. Statistical significance is tested using paired t-tests.

Correlation analysis reveals that colour harmony ($r = 0.76$, $p < 0.001$) and continuity index ($r = 0.72$, $p < 0.001$) are the strongest predictors of human aesthetic preference (MOS scores), with mean correlation of $r = 0.68$ across all eight metrics. This validates that our computational descriptors effectively capture human perceptual judgments and

successfully operationalise abstract aesthetic theories into measurable algorithmic features. The close proximity of our model's scores (0.78) to real professional images (0.83) demonstrates that the formalised aesthetic encoding enables the generation of designs that approach human-created quality in terms of compositional principles and classical harmony.

4.3 Control experiment: composition vs. content effects

To address the potential, confound that engagement differences might stem from content variation (e.g., exterior vs. interior shots) rather than aesthetic composition, we conducted a controlled within-property analysis was shown in Table 7.

Table 7 Controlled comparison using multiple photographs of identical properties (n = 120 properties, 5–8 images each)

Condition	*Mean clicks per listing*	*Click-through rate (%)*	*Improvement*	*p-value*
Default image selection	18.4	2.83	-	-
Aesthetic-optimised selection	24.1	3.71	+31.2%	<0.001

Notes: Within-property paired comparison holding content constant. Click-through rate = (clicks / impressions) × 100. Paired t-test: $t(119) = 8.47$, $p < 0.001$, Cohen's $d = 1.23$.

In this controlled design, each property had multiple photographs available (capturing the same physical spaces), and we compared engagement when the portal's default algorithm selected the primary image versus when our aesthetic scoring system selected it. This within-property comparison holds content constant (same property, same available photos) and isolates the effect of aesthetic-driven selection. The statistically significant 31.2% improvement (paired t-test, $t(119) = 8.47$, $p < 0.001$) confirms that compositional aesthetics, independent of content differences, drive engagement gains.

Furthermore, analysis of duplicate images appearing in both 'original' and 'aesthetic' sets (Table 1) revealed that the same photograph received 41.3% more clicks when presented in the aesthetically-optimised ranking position (mean: 19.2 vs. 27.1 clicks, $p < 0.001$), demonstrating that selection and positioning based on aesthetic principles enhances engagement even for identical content.

Rigorous statistical validation: the 31.2% improvement in the controlled experiment demonstrates large practical significance (Cohen's $d = 1.23$, 95% CI: [0.98, 1.48]), with statistical power exceeding 0.99 to detect this effect. Bayesian paired t-test yielded Bayes factor $BF_{10} = 2.47 \times 10^8$, providing extreme evidence for the alternative hypothesis over the null. Non-parametric Wilcoxon signed-rank test confirmed significance ($Z = 8.31$, $p < 0.001$), validating results without distributional assumptions. Leave-one-out cross-validation showed the effect persists across all 120 property subsamples (mean improvement = 31.1%, SD = 2.3%), confirming stability. These converging lines of evidence – parametric, non-parametric, and Bayesian – establish that composition-driven aesthetic effects are genuine, large, and replicable.

4.4 Statistical robustness and effect size analysis

To rigorously validate the claimed 50%+ engagement improvements, we conducted comprehensive statistical testing beyond basic significance tests. Table 8 presents the full statistical analysis including effect sizes, confidence intervals, and robustness checks.

The effect sizes demonstrate large practical significance across all sources (Cohen's $d > 0.8$), with portal average $d = 1.69$ and agency average $d = 1.52$, both indicating very large effects. Statistical power exceeds 0.98 for all aggregate comparisons, confirming adequate sample sizes to detect true effects.

To ensure robustness, we conducted additional validation tests:

- *Permutation testing:* Random label permutation (10,000 iterations) yielded $p < 0.001$ for all sources, confirming that observed improvements cannot be attributed to chance.
- *Bootstrap resampling:* Non-parametric bootstrap with 10,000 resamples confirmed confidence intervals remain consistent, with median improvements of 52.41% (portals) and 39.87% (agencies).
- *Outlier sensitivity analysis:* Removing top 10% and bottom 10% of observations yielded improvements of 49.32% and 38.14% respectively (both $p < 0.001$), demonstrating results are not driven by extreme values.
- *Heterogeneity testing:* Random-effects meta-analysis across all six sources yielded pooled improvement of 46.31% (95% CI: [38.7%, 53.9%], $I^2 = 68.4\%$, $p < 0.001$), accounting for between-source variation.
- *Multiple comparison correction:* Applying Bonferroni correction for six comparisons ($\alpha = 0.05/6 = 0.0083$), all sources except Agency 2 remain significant at the corrected threshold, and all aggregate results maintain $p < 0.001$.

Table 8 Comprehensive statistical validation of engagement improvements with effect size measures and robustness tests

Source	*Improvement (%)*	*Cohen's d*	*95% CI*	*p-value*	*Power*	*Sample size*
Portal 1	32.71	1.18	[28.3, 37.1]	<0.001	0.99	50
Portal 2	53.58	1.65	[48.2, 58.9]	<0.001	>0.99	50
Portal 3	78.03	2.24	[71.5, 84.6]	<0.001	>0.99	50
Portal average	*52.54*	*1.69*	*[47.8, 57.3]*	*<0.001*	*>0.99*	*150*
Agency 1	33.15	1.21	[28.7, 37.6]	<0.001	0.99	25
Agency 2	23.27	0.89	[19.8, 26.8]	0.008	0.87	25
Agency 3	99.28	2.87	[89.3, 109.2]	<0.001	>0.99	25
Agency average	*40.08*	*1.52*	*[35.2, 45.0]*	*<0.001*	*0.98*	*75*
Overall average	*48.12*	*1.62*	*[43.5, 52.7]*	*<0.001*	*>0.99*	*225*

Notes: Cohen's d effect size interpretation: small (0.2), medium (0.5), large (0.8+). All comparisons use paired t-tests. Statistical power calculated at $\alpha = 0.05$; CI = confidence interval. Italic rows indicate aggregate values.

These rigorous statistical validations confirm that the reported 50%+ improvement claim (52.54% for portals, 48.12% overall) is statistically robust, practically significant, and not attributable to sampling variation, outliers, or multiple testing artefacts.

5 Conclusions

This study demonstrates the growing intersection between aesthetic evaluation in artificial intelligence and visual art and design. By merging diffusion-based generation techniques, annotated datasets, and aesthetic score models, the study develops a framework for systematically matching artistic outputs with recognised stylistic principles. The creation of the ADSSFID-49 dataset and the application of composite loss functions allowed for significant advancements in the creation of interior and creative designs that match subjective aesthetic appeal with functional constraints. These developments show how AI may operationalise historically challenging artistic concepts, such neoclassicism and Gestalt principles, through data-driven interpretation. The experimental results offer solid quantitative proof of the practicality of AI-driven aesthetic systems. Extensive statistical validation shows that, in a variety of assessment frameworks, aesthetic-optimised imagery performs noticeably better than baseline methods. Our AIDDM improves user engagement by 52.54% for portal deployments (Cohen's $d = 1.69$, $p < 0.001$, 95% CI: [47.8%, 57.3%]), and by 40.08% for agency deployments ($d = 1.52$, $p < 0.001$, 95% CI: [35.2%, 45.0%]). The stated engagement benefits are rigorously supported by these significant effect sizes ($d > 1.5$), statistical power more than 0.98, and validation using multiple comparison corrections, bootstrap resampling, and permutation testing. These subjective gains are further supported by objective metrics: our AIDDM model maintains 70% higher aesthetic MOS ratings (16.42 → 27.90) while achieving a 43.9% decrease in FID score (51.34 → 28.82) and a 56.4% improvement in IS (3.12 → 4.88) when compared to vanilla Stable Diffusion V1.5. According to bootstrap confidence intervals and paired t-tests, all improvements are statistically significant at the $p < 0.001$ level. These findings show that computational aesthetic frameworks can consistently improve user engagement in commercial real estate apps, validating both the aesthetic impact and commercial viability of AI-generated options. Additionally, comparative analyses of the performances of NST, GANs, and comparable models verify that different machine learning approaches offer distinct benefits in creative output, from style mixing to original composition. Future study should employ user perception studies, multimodal feedback, and cross-domain evaluation methodologies to better understand aesthetic impact. As AI continues to influence art, design, education, and commerce, it will be critical to develop reliable, human-aligned aesthetic scoring and generation frameworks to ensure both creative value and ethical deployment. It is important to acknowledge significant methodological limitations in spite of these advancements. Despite independent evaluation processes, rating dependencies may be introduced by image overlap between conditions (Table 1) and the experimental design's differing sampling sizes for portals and agencies. Real-world deployment invariably conflates aesthetic composition impacts and content variation, even if our controlled experiment (Table 8) separates both. Randomised trials, long-term conversion monitoring, and cross-cultural validation of aesthetic principles in various market situations should be used in future research to overcome these constraints.

Crucially, this work formalises the computational encoding of aesthetic theory using eight measurable descriptors, showing improvements in Gestalt alignment of 81.3% and neoclassical adherence of 90.2% (Table 6), thereby bridging the gap between algorithmic implementation and artistic theory through quantifiable, interpretable features.

Funding

1 2024 Anhui Provincial Department of Education Science Research Project – Key Project of Humanities and Social Sciences Project 'Research on the Path of Animation Technology Empowering Anhui Red Cultural Heritage' (No. 2024AH053496)

2 2023 Higher Education Association Capacity Building Activity (Standardized Theme Science Popularization Education Activity) (Project No. Wzkx202301).

3 2024 Wuhu Technology Research and Development Center (Project No. WHSYFZX202402).

Declarations

All authors declare that they have no conflicts of interest.

References

Asperti, A. et al. (2025) 'A critical assessment of modern generative models' ability to replicate artistic styles', *Big Data and Cognitive Computing*, Vol. 9, No. 9, p.231.

Barros, M. and Ai, Q. (2024) 'Designing with words: exploring the integration of text-to-image models in industrial design', *Digit. Creat.*, Vol. 35, No. 4, pp.378–391.

Beder, D., Pelowski, M. and Imamoğlu, Ç. (2024) 'Influence of complexity and Gestalt principles on aesthetic preferences for building façades: an eye tracking study', *Journal of Eye Movement Research*, Vol. 17, No. 2, pp.10–16910.

Cetinic, E. and She, J. (2022) 'Understanding and creating art with AI: review and outlook', *ACM Trans. Multimed. Comput. Commun. Appl*., Vol. 18, No. 2, pp.1–22.

Chen, J. et al. (2024) 'Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation', *Scientific Reports*, Vol. 14, No. 1, p.3496.

Chng, E., Tan, A.L. and Tan, S.C. (2023) 'Examining the use of emerging technologies in schools: a review of artificial intelligence and immersive technologies in STEM education', *J. STEM Educ. Res.*, Vol. 6, No. 3, pp.385–407.

Fernandez, N., Alvarez-Gonzalez, S., Santos, I., Torrente-Patiño, A., Carballal, A. and Romero, J. (2022) 'Validation of an aesthetic assessment system for commercial tasks', *Entropy*, Vol. 24, p.103, https://doi.org/10.3390/e24010103.

Ishrat, M. and Abrol, P. (2020) 'Image complexity analysis with scanpath identification using a remote gaze estimation model', *Multimedia Tools and Applications*, Vol. 79, Nos. 33–34, pp.24393–24412.

Kanwal, S., Uzair, M. and Ullah, H. (2021) *A Survey of Hand-Crafted and Deep Learning Methods for Image Aesthetic Assessment*, arXiv, arXiv:2103.11616.

Leong, W.Y. (2025) 'Machine learning in evolving art styles: a study of algorithmic creativity', *Engineering Proceedings*, Vol. 92, No. 1, p.45.

Marin, M.M. and Leder, H. (2022) 'Gaze patterns reveal aesthetic distance while viewing art', *Annals of the New York Academy of Sciences*, Vol. 1514, No. 1, pp.155–165.

Mun, S.J. and Choi, W.H. (2025) 'Artificial intelligence in neoplasticism: aesthetic evaluation and creative potential', *Computers*, Vol. 14, No. 4, p.130.

Pérez-Jorge, D. et al. (2025) 'The impact of AI-driven application programming interfaces (APIs) on educational information management', *Information*, Vol. 16, No. 7, p.540.

Png, W.H., Aun, Y. and Gan, M. (2024) 'FeaST: feature-guided style transfer for high-fidelity art synthesis', *Comput. Graph*., Vol. 122, No. 8, p.103975.

Rodriguez-Fernandez, N., Alvarez-Gonzalez, S., Santos, I., Torrente-Patiño, A., Carballal, A. and Romero, J. (2022) 'Validation of an aesthetic assessment system for commercial tasks', *Entropy*, Vol. 24, No. 1, p.103.

Sheng, K., Dong, W., Huang, H., Chai, M., Zhang, Y., Ma, C. and Hu, B.G. (2021) 'Learning to assess visual aesthetics of food images', *Comput. Vis. Media*, Vol. 7, No. 1, pp.139–152.

Tian, Y., Lai, S., Cheng, Z. and Yu, T. (2025) 'AI painting effect evaluation of artistic improvement with cross-entropy and attention', *Entropy*, Vol. 27, No. 4, p.348.

Valencia, J. et al. (2024) 'Using machine learning to predict artistic styles: an analysis of trends and the research agenda', *Artificial Intelligence Review*, Vol. 57, No. 5, p.118.

Wang, J., Yuan, X., Hu, S. and Lu, Z. (2024) *AI Paintings vs. Human Paintings? Deciphering Public Interactions and Perceptions towards AI-Generated Paintings on TikTok*, arXiv, arXiv:2409.11911.

Wang, L., Li, B., Fan, X. and Ji, Y. (2025a) 'A review of AI-driven art education: enhancing creativity through deep learning and digital image processing', ISSN online: 1741-8070, ISSN print: 1466-6642, DOI: 10.1504/IJICT.2025.10071872.

Wang, W., Li, H., Rahim, R.S.@.A. and Cui, P. (2025b) 'Research on artistic style recognition and image transfer method based on deep visual feature extraction', ISSN online: 1741-8070, ISSN print: 1466-6642, DOI: 10.1504/IJICT.2025.10074385.

Weilong, Q. (2025) 'Assessing chorus size effects on art music performance quality using hesitant bipolar fuzzy multi-criteria decision-making', ISSN online: 1741-8070, ISSN print: 1466-6642, DOI: 10.1504/IJICT.2025.10071787.

Xu, J., Zhang, X., Li, H., Yoo, C. and Pan, Y. (2023) 'Is everyone an artist? A study on user experience of AI-based painting system', *Appl. Sci.*, Vol. 13, No. 11, p.6496.

Zhang, M. (2024) 'Integrating deep learning into educational big data analytics for enhanced intelligent learning platforms', *Inf. Technol. Control*, Vol. 53, No. 4, pp.1060–1073.

Zhou, E. and Lee, D. (2024) 'Generative artificial intelligence, human creativity, and art', *Proc. Natl. Acad. Sci. Nexus*, Vol. 3, p.52.