Bachelor Computer Science

Public Perception of AI Creativity

with Verbalized Sampling

Tim Boukens

Supervisors:
Rob Saunders
Tessa Verhoef

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

1/07/2026

Abstract

Large language models (LLMs) have raised a new question: Can AI models be creative? Mode collapse occurs frequently: the tendency of these models to produce boring outputs that humans tend to prefer. Verbalized Sampling (VS) [ZYC+25] is a recently proposed prompting technique that counteracts mode collapse: a model is instructed to generate several candidate descriptions. Together with self-estimated probabilities of occurrence, the least probable candidate is selected. This thesis investigates whether images generated via VS are perceived as more creative than images generated with standard prompting. It also looks at the effect of the level of transparency on the creativity ratings. A between-subjects online experiment assigns participants to one of three transparency conditions :no information about VS, the VS method disclosed, or the VS mechanism explained. It then collects Likert-scale ratings on four creativity sub-dimensions: Creativity, Novelty, Value, and Surprise. The data is then analysed using Cronbach's α for scale validation followed by a mixed ANOVA. Preliminary results show that VS images are rated significantly higher than standard images on Creativity, Novelty, and Surprise regardless of transparency condition, while transparency itself raises ratings broadly. It does however not solely raise VS-image creativity ratings.

Contents

1 Introduction

The question of whether a machine can be creative predates the machines themselves. Ada Lovelace argued that a computing engine could only ever do what it was explicitly told to do, and that any seemingly inventive output should be credited to the engineer who programmed it, not to the machine [Bod04]. Almost two centuries later, that objection has resurfaced in a very concrete form. AI-generated images and text have gained substantial public attention and commercial acceptance over the past few years: models like DALL-E, Stable Diffusion, and Midjourney are frequently used by millions of people for tasks ranging from illustration to advertising [EHT23]. Their slowly growing acceptance, however, does not settle the question Lovelace raised. Two distinct issues are at stake. The first is a question about the generative process itself: can a language model be prompted or tuned to produce outputs that are more diverse and therefore more creative, rather than defaulting to safe, repetitive responses? The second is a question about perception: once such an output exists, does knowing how it was produced change how creative people judge it to be? This thesis keeps these two questions distinct throughout, since conflating them risks obscuring which of the two is actually responsible for any change in creativity ratings.

Research on computational creativity has long documented a human bias on the perception side: people tend to rate the same work lower when told it was produced by a machine, even when the work itself is identical to a human-made counterpart [CMSW18]. On the generative side, changing how prompts are constructed has been explored as a way to counteract the tendency of language models to default to safe, repetitive outputs. Verbalized Sampling (VS) [ZYC+25] is one such advance: a language model is prompted to enumerate candidate outputs together with self-estimated probabilities, and the least probable candidate is selected. VS encourages outputs that are almost never selected under standard conditions, potentially leading to more diverse and thus more creative outputs. Zhang et al. [ZYC+25] validate this approach on text generation; this thesis instead applies VS to the generation of prompts for image generation, an application that, to my knowledge, has not previously been explored in the literature. This constitutes a methodological contribution in its own right, independent of the perception question the thesis goes on to ask.

VS-generated images may thus be surprising and unexpected. But if humans already tend to dislike AI-generated images, can the improved diversity of VS-generated images, together with transparency about how they were generated, shift those judgements?

1.1 Research Question

This thesis addresses the following two research questions, corresponding to the generative and perceptual issues introduced above:

1. Does the use of Verbalized Sampling to generate elaborated prompts increase the perceived creativity of the resulting images, compared to standard prompting?
2. Does revealing the use of Verbalized Sampling in the image generation process further affect how creative those images are perceived to be?

The questions are explored through a three-condition between-subjects experiment in which both image generation methods (standard and VS) are shown to every participant. The experiment also varies the degree of information disclosed to participants between conditions (none, method name only, or a full explanation of the generation process). Participants evaluate the same 10 standard images in all conditions, while the level of transparency attached to the 10 VS-generated images varies between conditions.

1.2 Contributions

This thesis makes the following contributions.

- An empirical investigation of human creativity judgements for VS-generated versus standard AI images, using a multi-variable creativity rating.
- A novel application of Verbalized Sampling regarding the generation of elaborated image-generation prompts, rather than plain text directly. This also includes evidence that this approach produces images rated as more diverse and creative than standard prompting.
- A manipulation of transparency (three levels) to test whether disclosure of the generation process changes creativity attributions.

1.3 Thesis Overview

The remaining sections are covered as follows: Section 2 provides background on AI creativity, mode collapse, and verbalized sampling. Section 3 reviews related work on human perception of computational creativity. Section 4 describes the experimental design, materials, participants,

and analysis plan. Section 5 reports preliminary results. Section 7 describes the main conclusion and gives directions for future work.

2 Background and Key Definitions

2.1 Creativity and Its Dimensions

Boden [Bod04] compares three creativity parameters. Exploratory creativity entails exploring the limits of a conceptual space. Another frequent type is combinational(novel combinations of familiar ideas). Transformational creativity means altering the space itself. Boden's notion of this conceptual space, a structured way of exploring or reshaping the creative process, has since been formalised more rigorously by Gärdenfors [Gär00]. Researchers in machine learning have drawn an explicit parallel between conceptual spaces and the latent or embedding spaces learned by neural networks [Var19]. Under this view, exploratory creativity can be understood as movement through such a space. A creative output can be seen as one that occupies a remote or under-explored region of the space while remaining recognisably related to a known concept. This framing motivates the approach taken in this thesis: Verbalized Sampling, introduced in Section 2.3, can be understood as a mechanism for deliberately steering a language model's output towards such under-explored regions of its own output distribution.

For measurement purposes, creativity researchers commonly decompose creativity into sub-dimensions reflecting different aspects of what makes an output creative. Following this convention, this study treats creativity as a composite of four constructs: Creativity, Novelty, Value, and Surprise. The precise methods of these constructs as measurable rating items is deferred to the methodology (Section 4.6). Here, the constructs are introduced only at the conceptual level needed to motivate the experimental design.

2.2 Mode Collapse

Mode collapse is not unique to large language models. The term originates in the literature on Generative Adversarial Networks (GANs), where a generator learns to produce a tiny portion of the available target data distribution. This is usually the case in order to reliably fool the discriminator, rather than capturing the full diversity of the training data [GPAM+14, TTT20]. Considerable effort in the GAN literature has gone into mitigating this failure mode. For example, through modified objective functions, minibatch discrimination, or unrolled

optimisation [TTT20].

Mode collapse in large language models arises differently, but with a similar end result. LLMs are commonly adapted by using reinforcement learning based on human feedback (RLHF). This favors output responses that human raters prefer [OWJ+22]. Because human raters tend to favour safe, familiar, and expected answers, this training creates a typicality bias: the model learns that high-probability, conventional responses are systematically rewarded over unusual ones. The result is that the model's output distribution collapses onto a narrow band of high-probability responses, even though the model retains the capacity to generate a much wider range of plausible outputs [ZYC+25]. This is particularly problematic for creative applications of LLMs, where the desirable outputs are precisely those that are unusual, unexpected, or rare. These same outputs are surpressed by RLHF-style training.

A number of approaches besides Verbalized Sampling have been proposed to counteract mode collapse in language models. This includes raising the sampling temperature, nucleus or top-p sampling, and prompting-based diversity techniques that explicitly ask the model to avoid its most typical response [ZYC+25]. These approaches generally trade off fluency or coherence against diversity, or rely on the model's surface-level output token probabilities rather than its self-assessed probability of an entire response. The following subsection introduces Verbalized Sampling, the specific approach explored in this thesis.

2.3 Verbalized Sampling

Zhang et al. [ZYC+25] proposed Verbalized Sampling (VS) as a prompting-based way to counteract mode collapse without retraining the model. Rather than sampling directly from the model's token distribution, the model is explicitly prompted to produce a set of n candidate responses to a single underlying request, along with a self-estimated probability for each candidate. This can be done by prompting the model with an instruction such as "generate five possible responses to the following prompt, along with the probability that each would typically be generated". The candidate the model assigns the lowest self-estimated probability is then selected as the final output. By selecting from the tail of the model's own estimated distribution rather than its mode, VS encourages outputs that are statistically less likely to be produced under standard sampling, while still being coherent responses to the original request.

Zhang et al. [ZYC+25] validate VS primarily on text generation tasks such as creative writing and dialogue, showing that it substantially increases output diversity compared to standard

sampling strategies. Since the original publication, a small number of works have begun to apply or extend VS to other settings; at the time of writing, however, no published work applies VS to the generation of prompts for image generation models. The application explored in this thesis is therefore treated here as a novel methodological contribution rather than a direct replication of prior work, and is further outlined in Section 4.3.

2.4 AI Image Generation and Diffusion Models

Modern text-to-image generation has evolved considerably in how prompts are constructed. Early systems built around CLIP-guided generation, such as early Midjourney versions, were typically driven by dense, comma-separated keyword lists. These were informally known as "prompt word clouds". These were used instead of grammatical natural language, since CLIP-based guidance responded primarily to the presence of salient tokens rather than syntactic structure [RKH+21]. The shift towards multi-modal models with stronger native language understanding, such as ChatGPT-integrated image generation, has made fluent natural-language prompting increasingly common and effective. These models can parse descriptive, compositional instructions rather than relying on keyword density. This evolution is relevant to the mode collapse problem discussed above: natural-language prompting interfaces make it easier to elicit a wide range of phrasings for a given idea, which is precisely the property that Verbalized Sampling exploits when generating elaborated prompts.

The images used in this study are generated by a text-to-image image generation model. These so called diffusion models operate by learning to decode images with added noise back to their training versions. In the next stage, a random noise tensor is denoised step-by-step, guided by a text embedding derived from the input prompt [RBL+22]. The same base prompt (e.g. "a bear in the forest") is used for both the standard and VS conditions; only the elaborated prompt differs between conditions.

3 Related Work

While Section 2 concerned the generative side of this thesis, this section turns to the perceptual side: how the creativity of computational outputs, once produced, is judged by human observers. This question has a long history. Ada Lovelace's nineteenth-century objection that a machine's apparently inventive output should be credited to its programmer rather than the machine itself is widely regarded, following Boden's discussion of it [Bod04], as an early statement of

the scepticism that still shapes how people evaluate AI-generated creative work today.

3.1 Types and Dimensions of Creativity

Creativity researchers have proposed several complementary lenses for analysing creative work. Rhodes [Rho61] introduced what is now commonly known as the Four-P model. This model seperates Process, Product, Person and Press (the environment) as four distinct aspects of any creative episode. This distinction between the *process* by which a creative artefact is produced and the *product* itself is central to the present thesis, since the experiment manipulates knowledge of the process (via the transparency conditions) while holding the product (the images themselves) fixed. Jordanous [Jor16] translates the Four-P framework specifically into the context of computational creativity, making it directly relevant to evaluating AI-generated work.

A second distinction concerns the *scale* of creativity. Researchers commonly distinguish Big-C creativity, the historically significant creative achievements of a field (e.g. general relativity), from little-c creativity, the day to day expressions of ordinary people. Kaufman and Beghetto [KB09] extend this into a Four-C model by adding Pro-C (professional-level) and mini-C (personal, everyday) creativity. The images evaluated in this thesis sit closer to the mini-c and little-c end of this spectrum, with the rating task potentially also engaging Pro-C judgements among more experienced raters. Glaveanu and Kaufman [GK19] synthesise the Four-P and Four-C frameworks into a single matrix, mapping how different lenses on creativity interact across different scales, and provide a useful integrative perspective for situating the present study within the broader creativity literature.

3.2 Human Bias Against Computational Creativity

Chamberlain et al. [CMSW18] asked participants to rate computer-generated artworks on aesthetic and creative dimensions, holding the artworks themselves constant across conditions. They found that participants rated the same work lower on creativity when told it was computer-generated than when no such information was given, suggesting a preference for human-attributed creative work over machine-attributed work of identical quality. Mumford and Ventura [MV15] report a related form of scepticism toward creative computing systems more broadly, finding that audiences are often reluctant to accept a computational system as genuinely creative even when its outputs are evaluated favourably, which is directly relevant to the question of whether disclosing a method like VS can shift such scepticism.

A comparable bias has been documented outside the visual domain. In music, Moffat and Kelly [MK06] found that listeners systematically rated human-composed pieces more favourably than identical pieces attributed to an algorithm. Pasquier et al. [PBT+16] report the same effect, though they find it weaker among participants with greater musical expertise, suggesting that domain familiarity may moderate the bias against computational creativity.

3.3 The Role of Process Knowledge

Beyond the binary question of human versus machine attribution, prior information about *how* a creative process unfolded can shape evaluations of its output. Massi et al. [MVD20] found that consumers incorporate descriptions of a production process into their judgements of the resulting product, with descriptions that emphasise effort and deliberate work generally producing more favourable evaluations. Agudo et al. [ALM22] report an analogous effect specifically for visual art: identical artworks were judged differently depending on whether participants believed they had been produced by a human or by an AI system, indicating that beliefs about the creative process, independent of the artefact itself, influence perceived value. Taken together, these findings motivate the prediction that explaining how VS deliberately selects low-probability candidates could lead viewers to perceive the underlying process as more inventive, and thereby rate the resulting images as more creative. That is the central hypothesis tested by the transparency manipulation in this thesis.

3.4 Verbalized Sampling and Diversity

Zhang et al. [ZYC+25] demonstrate that VS significantly increases the diversity of LLM outputs compared to standard sampling strategies such as highest-probability decoding. However, their validation is conducted almost entirely on text generation tasks, and does not measure human perception of the resulting outputs in terms of creativity, nor does it extend to image generation. This leaves open both the question of whether VS-induced diversity translates into perceived creativity in a different modality, and the question of whether transparency about VS specifically (rather than AI involvement in general) affects that perception.

3.5 Transparency and AI Perception

A separate line of work examines how transparency about a system's involvement affects user trust and evaluation more generally. In creative contexts, simply disclosing that AI was involved tends to reduce perceived creativity relative to no disclosure. This reduction can, however, be offset by positively framed descriptions of the technique involved, as Chamberlain et al. [CMSW18] shows. This returns to the same paradigm discussed earlier, this time manipulating how the computational process is described rather than merely whether it is disclosed. What remains unexplored is whether this moderating effect of framing extends to a specific prompting technique like VS. This in comparison to generic AI disclosure. That is the gap this thesis aims to address. Related work on *framing* in computational creativity more broadly, including Colton's writing on how systems can be framed to an audience to shape perceptions of their creativity [CW12], offers a complementary perspective on this question.

4 Methodology

This section outlines the image stimuli generation and the experiment used to evaluate them. The VS image generation pipeline is introduced first, since the design of the experiment depends on understanding what distinguishes a VS-generated image from a standard one. The experimental design, stimuli, participants, procedure, rating instrument, and analysis plan then follow in order.

4.1 VS Image Generation Pipeline

To apply Verbalized Sampling to image generation, a two-stage pipeline was developed. In the first stage, a language model is used to elaborate a brief input prompt into a richer, less typical description. In the second stage, that elaborated description is passed to a diffusion model to produce the final image. This pipeline is illustrated with a worked example below.

Stage 1: Verbalized Sampling over prompt descriptions Given a brief input prompt (e.g. "a bear in the forest"), the language model is instructed as follows:

> *Generate five possible elaborated image descriptions for the following prompt, along with the probability (as a decimal between 0 and 1) that each description is the one you would typically generate. List them from most probable to least probable.*

The model returns five candidate descriptions together with self-estimated probabilities. The candidate assigned the lowest probability—provided it is at or below $p \leq 0.10$—is selected as the elaborated prompt. If no candidate meets this threshold, the process is repeated. Because the goal is to test the effect of VS rather than editorial preference over outputs, the first image produced from the selected prompt was always accepted without filtering.

Stage 2: Image generation from the elaborated prompt The selected elaborated prompt is passed image model (the same model used for the standard condition), which produces the final image.

For the input prompt "a bear in the forest", the language model returned the following five candidates (probabilities as self-estimated by the model):

1. A grizzly bear rendered entirely in iridescent soap-bubble geometry, translucent prismatic surfaces, floating in a void, macro photography lightings $p = 0.09$
2. A medieval illuminated manuscript depicting a bear as a heraldic creature, gold leaf borders, Latin marginalia, aged vellum texture, top-down flat lay $p = 0.04$
3. A bear skeleton posed upright in a Victorian natural history museum diorama, gas-lamp lighting, mahogany display case, hand-written specimen cards $p = 0.06$
4. A polar bear seen from directly above, floating on a single sheet of translucent blue ice, drone perspective, minimal negative space, National Geographic editorial style $p = 0.05$
5. **A Scandinavian folk-art bear in the style of a Dala horse, carved wood texture, flat bright reds and blues, white floral motifs, studio white background** $p = 0.03$

Candidate 5 was selected ($p = 0.03 \leq 0.10$). The standard image was generated directly from "a bear in the forest"; the VS image was generated from the selected candidate. Figure 1 shows the two resulting images side by side.

4.2 Experimental Design

The experiment uses a 3 (transparency condition) $\times$ 2 (image type) mixed design. Transparency condition is the between-subjects factor; image type (standard vs. VS) is the within-subjects factor, since every participant rates both types. The central purpose of the design is to disentangle two effects: whether VS images are judged as more creative regardless of what participants know about how they were made (image type main effect), and whether knowledge of the generation process selectively amplifies those judgements (condition $\times$ image type interaction).

Figure 1: Standard image (left) generated from "a bear in the forest" and VS image (right) generated from the Dala-horse elaboration ($p = 0.03$).

All participants see the same ten standard images and the same ten VS images. What varies between conditions is the process information provided alongside the VS images. Standard images are always accompanied by the bare input prompt only, serving as the within-session baseline for each participant. The three conditions differ in the information provided for VS images.

With condition 1 (bare prompt), VS images are presented with the bare input prompt only, identical to the information given for standard images. This condition provides the cleanest comparison of perceived creativity between image types, since participants receive no information about the generationmethod and cannot distinguish VS images from standard images on that basis. Condition 2 (elaborated prompt) expands this by additionally labelling the images with the elaborated prompt that was used to generate them (e.g. "A Scandinavian folk-art bear in the style of a Dala horse…"). Participants can see what the model produced as a description, but receive no explanation of how that description was generated. Condition 3 (full VS explanation) finish this process with a label with the elaborated prompt and a one-sentence explanation of the Verbalized Sampling method, including the selected probability value. This is the maximum process information condition.

The conditions are summarised in Table 1. Participants are, in a random order, added to one of the three conditions at the start of the survey, with each condition assigned with equal probability ($\frac{1}{3}$).

Figure 2 shows representative screenshots of the image presentation interface for each condition, illustrating concretely what information participants saw alongside each VS image.

Condition	Images	Process information provided for VS images
1	Standard (10) + VS (10)	Bare input prompt only (same as standard images)
2	Standard (10) + VS (10)	Bare input prompt + elaborated prompt used for generation
3	Standard (10) + VS (10)	Bare input prompt + elaborated prompt + one-sentence VS method explanation and selected probability

Table 1: The three experimental conditions. All participants rate all 20 images. Standard images are shown with the bare input prompt in every condition. Each condition increases the amount of process information provided for VS images.

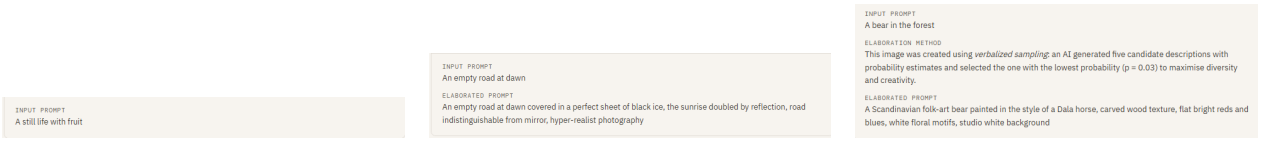

Figure 2: Survey interface for Condition 1 (left), Condition 2 (centre), and Condition 3 (right), showing the process information presented alongside a VS image in each case. Condition 2 shows only the bare input prompt. Condition 3 additionally shows the elaborated prompt. Condition 3 additionally shows a one-sentence VS mechanism explanation and the selected probability.

4.3 Stimuli

Ten input prompts were chosen to represent a range of common image subjects while remaining brief and genre-neutral: “a bear in the forest”, “a city at night”, “a boat on water”, “a mountain landscape”, “a figure in a doorway”, “a still life with fruit”, “an empty road at dawn”, “a garden in bloom”, “a child looking upward”, and “a bird in flight”. For each prompt, one standard image and one VS image were produced using the pipeline described in Section 4.1, yielding ten matched pairs.

Standard images: Each input prompt was passed directly to the diffusion model without elaboration. Figure 5 shows all ten standard images; larger versions are included in Appendix A.

Figure 3: The ten standard images, generated directly from the input prompts (left to right, top row: bear, city, boat, mountain, doorway; bottom row: fruit, road, garden, child, bird). Larger versions are shown in Appendix A.

VS images: For each input prompt, the VS pipeline (Section 4.1) produced an elaborated prompt with $p \leq 0.10$; this prompt was then used to generate the VS image. Table 2 lists all elaborated prompts and their associated probability estimates, keyed by a short identifier matching the input prompt. Figure 6 shows all ten VS images; larger versions are included in Appendix A.

ID	Input prompt	VS elaborated prompt (selected candidate)	p
bear	a bear in the forest	Scandinavian folk-art bear in the style of a Dala horse, carved wood texture, flat bright reds and blues, white floral motifs, studio white background	0.03
city	a city at night	City at night from the International Space Station, curved Earth horizon, city grid as a glowing neural network, deep black space above	0.07
boat	a boat on water	Venetian gondola mid-crossing in heavy fog, gondolier as silhouette, water still as mercury, monochrome	0.03
mountain	a mountain landscape	Topographic contour map of a mountain range come to life in 3-D; paper-white terrain, ink contour lines raised as physical ridges, isometric view	0.06
doorway	a figure in a doorway	Figure in a doorway between two realities—snowy pine forest on one side, scorched desert on the other; figure itself motion-blurred	0.06
fruit	a still life with fruit	Dutch Golden Age still life recreated with fruit made of polished semiprecious stones—malachite grapes, carnelian pomegranate—on black velvet	0.02
road	an empty road at dawn	Empty road at dawn covered in a perfect sheet of black ice; sunrise doubled by reflection; road indistinguishable from mirror; hyper-realist photography	0.03
garden	a garden in bloom	Garden in bloom photographed entirely in ultraviolet light; petals reveal hidden nectar guides as vivid UV patterns; bees hovering	0.06
child	a child looking upward	Child looking upward at the Milky Way reflected in a flooded salt flat, standing at the precise centre, mirror symmetry perfect, Bolivia Salar de Uyuni	0.04
bird	a bird in flight	Bird in flight depicted as a single unbroken neon tube sculpture, bent into the form of outstretched wings, mounted on a dark gallery wall, pink glow	0.03

Table 2: Input prompts, VS elaborated prompts, and selected probabilities for each of the ten image pairs. Short IDs (bear, city, ...) are used to refer to prompt pairs throughout.

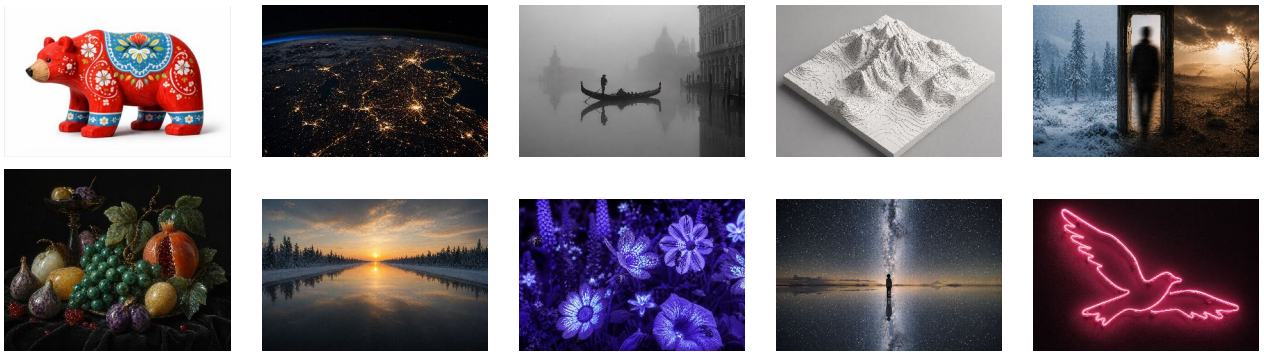

Figure 4: The ten VS images, each generated from the elaborated prompt selected by the VS pipeline (same order as Figure 5). Larger versions are shown in Appendix A.

4.4 Participants

The target sample size was approximately 30 participants, distributed across the three conditions (approximately 10 per condition). Participants were recruited via SurveySwap.io, social media, and academic networks to reduce selection bias. Inclusion criteria: 18 years or older, proficiency in English sufficient to read the rating items, completion of the attention check. Participants self-report their age range and familiarity with AI image generation (1 = not at all, 5 = extremely familiar) on the demographics page.

4.5 Procedure

The study was delivered as a single-page web application hosted at `ai-image-survey.netlify.app`. After reading an information sheet and providing demographic information, participants completed an attention check: they were explicitly instructed to select "3 — Neither agree nor disagree" on a Likert item. This check is a standard technique for identifying inattentive or random responders in online surveys [OMD09]; participants who failed it were excluded from analysis.

Participants then viewed 20 images presented in a random order (shuffled independently for each participant). For each image, they rated eight statements on a Likert scale with 5 points. That ranged each time from 1-strongly disagree, to 5-agree. This means using the rating instrument described in Section 4.6. An optional free-text comment box followed each image.

4.6 Rating Instrument

The four creativity constructs introduced in Section 2 (Creativity, Novelty, Value, and Surprise) were each operationalised using two Likert items, for eight items in total. The operationalisation is grounded in the sub-dimensional structure of each construct as identified in the creativity literature.

Creativity (two items) is measured at both the process and the product level, following Rhodes's [Rho61] distinction between the creative process and the creative product: one item asks about the perceived inventiveness of the process that produced the image; the other asks about the creativity of the resulting output.

Novelty (two items) is decomposed into statistical rarity and semantic novelty, reflecting the standard two-component definition of novelty [Bod04]: one item asks whether the participant has rarely seen something similar (statistical rarity); the other asks whether the image combines ideas in an unexpected way (semantic novelty, i.e. the unusualness of the concept rather than just its surface appearance).

Value (two items) is measured across its two most common sub-dimensions [Bod04]: aesthetic appeal (whether the image is visually pleasing) and functional usefulness (whether it could serve a purpose in a real context). These two aspects of value were kept as separate items rather than collapsed into a single rating, in order to test whether aesthetic and functional judgements behave similarly.

Surprise (two items) captures two related but distinct aspects of the unexpectedness of an image: the degree to which the image pleasantly surprised the participant, and the degree to which it turned out differently from what they expected. The distinction matters because pleasant surprise implies an positive reaction, while expectation violation is more neutral and concerns the distance between the image and what a standard prompt might be expected to produce.

The eight rating items are listed in Table 3.

4.7 Data Analysis Plan

Stage 1: Scale validation. Internal consistency within each construct is assessed using Cronbach's α. For each construct, if $\alpha \geq 0.70$, the two items are sufficiently correlated to be averaged into a single composite score; if $\alpha < 0.70$, the two items are retained as separate dependent variables. This stage therefore determines both the number of composite scores

Construct	Item ID	Statement
Creativity	creative_proc	The process used to make this image seems inventive.
	creative_out	The result of this image is creative.
Novelty	novelty_rare	I have rarely seen something like this.
	novelty_sem	This image combines ideas in an unexpected way.
Value	value_aesth	This image is visually pleasing.
	value_func	This image could be useful in a real context.
Surprise	surprise_unexp	This image pleasantly surprised me.
	surprise_viol	This image turned out differently from what I expected.

Table 3: Rating instrument: eight items across four constructs.

and the total number of dependent variables entering Stage 3.

Stage 2: Composite scores. For each construct that passes the Stage 1 threshold, the two items are averaged to yield one composite score per image per participant. Constructs that fail are carried forward as two separate scores. The set of dependent variables produced here, whether four composites or some mix of composites and individual items, is analysed in Stage 3.

Stage 3: Main analysis. A mixed ANOVA is run for each dependent variable from Stage 2, with image type (standard vs. VS) as a within subjects factor and transparency condition (2, 3, 4) as a between subjects factor. Key tests are the image type main effect (Research Question 1) and the condition $\times$ image type interaction (Research Question 2).

Bonferroni correction is applied across the full set of dependent variables produced by Stage 2 to control the type I error rate: the adjusted threshold is $\alpha = 0.05/k$, where k consists of the amount of dependent variables. Because k depends on the outcome of Stage 1, the exact threshold is determined after scale validation. AI familiarity and age range are included as covariates to assess whether they moderate the primary effects.

5 Results

5.1 Participants

Thirty participants completed the survey and passed the attention check. The sample was predominantly young adults: 14 participants (47%) were aged 18–24, 12 (40%) aged 25–34, 2 (7%) aged 45 or older, 1 aged 35–44, and 1 under 18. Self-reported AI image generation familiarity was moderate: 13 participants (43%) rated themselves at level 3 (moderately familiar), 10 (33%) at level 4 (very familiar), 5 (17%) at level 2 (slightly familiar), and 2 (7%) at level 5 (extremely familiar). After random assignment, 10 participants were allocated to each of the three conditions.

5.2 Stage 1: Scale Validation

Internal consistency was assessed with Cronbach's α for each of the four two-item sub-scales (Table 4).

Construct	**Cronbach's α**	**Result**
Creativity	0.890	Pass (≥ 0.70)
Novelty	0.827	Pass
Surprise	0.841	Pass
Value	0.681	Fail (< 0.70)

Table 4: Cronbach's α for each sub-scale. Creativity, Novelty, and Surprise pass the threshold and are averaged into composite scores. Value fails; its two items (*value_aesth* and *value_func*) are retained as separate dependent variables.

Three composites pass the threshold and are averaged into single scores per image per participant: $C_{\text{creativity}}$, C_{novelty}, and C_{surprise}. The two Value items are analysed separately, yielding five dependent variables in total. The Bonferroni-corrected significance threshold is adjusted to $\alpha = 0.05/5 = 0.010$ (rather than $0.05/8 = 0.00625$, which assumes all items remain independent from each other).

5.3 Stage 2: Composite Scores

Table 5 reports participant-level means and standard deviations for each composite and each dependent variable, broken down by image type and condition. Standard images serve as a

within-session baseline in the table; all participants rated the same ten standard images, so their scores are pooled across conditions.

VS images	Standard (baseline)				
Composite	*M*	*SD*	**Cond 1** *M*	**Cond 2** *M*	**Cond 3** *M*
Creativity	2.50	0.77	2.91	3.60	3.33
Novelty	2.15	0.66	2.71	3.19	3.04
Surprise	2.35	0.76	2.74	3.27	3.10
Value_aesth	3.16	0.84	3.04	3.73	3.37
Value_func	2.90	0.90	2.59	3.31	2.85

Table 5: Participant-level means and standard deviations for each composite score by image type and condition. Standard image scores are pooled across all 30 participants. VS image scores are reported separately per condition ($n = 10$ per condition).

5.4 Stage 3: Main Analysis

5.4.1 Assumption checks

Shapiro–Wilk tests indicated violations of normality in at least one group for Creativity ($p < .05$) and Value_aesth ($p < .05$). For these two variables, a ART-ANOVA [WFGH11] (non-parametric Aligned Ranks Transformation ANOVA) was conducted alongside the standard mixed ANOVA as a robustness check. Levene's test indicated no violations of homogeneity of variance.

5.4.2 Mixed ANOVA

A mixed ANOVA was conducted for each of the five dependent variables, with image type (standard vs. VS) as a withinsubjects factor and transparency condition (1, 2, 3) as a between subjects factor. Results are reported in Table 6.

5.4.3 ART-ANOVA robustness check

ART-ANOVA on Creativity and Value_aesth produced consistent results. For Creativity, the image type effect remained importantly significant ($F(1, 27) = 37.51$, $p < .001$) and the condition effect remained non-significant ($F(2, 27) = 2.59$, $p = .094$). For Value_aesth, the

Composite	Effect	F	df	p	η_p^2
Creativity	Condition	3.07	(2, 27)	.063	.185
	Image type	33.24	(1, 27)	$< .001^{***}$	.552
	Condition × Image type	0.17	(2, 27)	.847	.012
Novelty	Condition	2.63	(2, 27)	.091	.163
	Image type	41.24	(1, 27)	$< .001^{***}$	.604
	Condition × Image type	0.28	(2, 27)	.761	.020
Surprise	Condition	2.60	(2, 27)	.093	.161
	Image type	32.00	(1, 27)	$< .001^{***}$	.542
	Condition × Image type	0.53	(2, 27)	.594	.038
Value_aesth	Condition	3.51	(2, 27)	.044*	.206
	Image type	4.83	(1, 27)	.037*	.152
	Condition × Image type	0.42	(2, 27)	.664	.030
Value_func	Condition	3.46	(2, 27)	.046*	.204
	Image type	0.02	(1, 27)	.894	.001
	Condition × Image type	0.16	(2, 27)	.856	.011

Table 6: Mixed ANOVA results. Bonferroni-corrected threshold: $\alpha = 0.010$. *** $p < 0.001$; * $p < 0.05$ (nominally significant, does not survive Bonferroni correction). Effect sizes reported as partial η^2.

condition effect remained nominally significant ($F(2, 27) = 3.57$, $p = .042$), while the image type effect dropped to irrelevancy ($F(1, 27) = 3.49$, $p = .073$). No interaction was significant under ART-ANOVA for either variable. These results confirm that normality violations did not meaningfully distort the parametric findings.

5.4.4 Summary of findings

Three main findings emerge from the analysis.

VS images are consistently rated more creative than standard images. Image type produced a large, significant effect on Creativity ($\eta_p^2 = .552$), Novelty ($\eta_p^2 = .604$), and Surprise ($\eta_p^2 = .542$), all surviving Bonferroni correction ($p < .001$). VS images were rated higher than standard images across all conditions and all three composites.

Transparency condition raises ratings broadly, but not specifically for VS images. The condition main effect was nominally significant for Value_aesth and Value_func ($p = .044$ and $p = .046$, respectively), with Condition 2 producing the highest ratings. However, none of the condition effects survived Bonferroni correction, and the same pattern appeared in both standard and VS images within each condition, suggesting a session-wide mindset effect

rather than a targeted improvement in VS image perception.

No interaction between image type and condition. Crucially, no interaction effect was significant for any composite variable (all $p > .59$, all $\eta_p^2 < .04$). The advantage of VS images over standard images was approximately equal in size across all three transparency conditions. Transparency did not moderate the VS creativity premium.

6 Discussion

6.1 Interpretation

The central research question asks whether revealing a prompting method results in VS-generated art receiving higher creativity ratings than standard-prompted art. The results suggest a partial and nuanced answer.

VS images were reliably rated higher than standard images on Creativity, Novelty, and Surprise regardless of what information participants received. This confirms that verbalized sampling produces images that are perceptually distinct from standard-prompted images, and that human raters recognise this difference even without any explanation of the generation method. The effect is substantial, with partial η^2 values exceeding 0.54 across three composites, indicating that image type alone accounts for more than half the variance in perceived creativity.

However, transparency about the generation process did not specifically amplify the perceived creativity of VS images. The null interaction is the key result: the gap between VS and standard images was the same size in Condition 1 (no information), Condition 2 (elaborated prompt shown), and Condition 3 (full VS mechanism explained). Knowing *how* the image was made did not change *how much more* creative people judged it to be.

What transparency did appear to do is raise overall ratings across the session. Condition 3 participants gave higher scores to both their standard and VS images compared to other conditions, suggesting that reading an elaborate, detailed prompt primed a more engaged or generous evaluation mindset, rather than producing a targeted enhancement of VS image perception. This mirrors the findings of Massi et al. (2020) on process descriptions and perceived value: context shapes experience globally, not just locally.

Interestingly, Condition 3 — the only condition that explicitly explained the verbalized sampling mechanism — did not outperform Condition 2 on any composite. This suggests that

the additional information about how VS selects low-probability candidates did not further increase perceived creativity beyond simply showing participants the elaborated prompt itself. The image and its description appear to do most of the perceptual work; the underlying method adds little on top.

6.2 Instrument Reflection

The failure of the Value sub-scale to meet the internal consistency threshold (a = 0.7) warrants reflection beyond its analytic consequences. Two interpretations are possible. The first is that aesthetic appeal and functional usefulness are genuinely distinct psychological constructs, and that participants discriminated between them meaningfully: an image can be visually striking without being practically applicable, and vice versa. Under this reading, the low a is not a flaw but a finding . Value might not be a unitary dimension in the context of AI-generated images, and treating it as one would have obscured real variation. This reading is supported by the descriptive data: Value_aesth means were notably higher than Value_func means across all conditions, suggesting participants did not treat the two items as interchangeable. The second interpretation is that the two items were simply poorly matched in their surface phrasing or difficulty, producing inconsistent responses not because participants held different views on aesthetics versus function, but because the items failed to jointly operationalise a coherent construct. Distinguishing between these two explanations would require either a richer set of value items, perhaps four or six rather than two , or qualitative follow-up to establish how participants understood the distinction. Future work should address this directly, since value is arguably the dimension most likely to differ between standard and VS images in applied contexts where both creativity and practical usefulness matter.

6.3 Limitations

Several limitations constrain the interpretation of these results.

With $n = 10$ per condition, the study is underpowered to detect small-to-medium effects reliably. The condition effects on Creativity, Novelty, and Surprise all fell in the range $p = .06$–$.09$, which may reflect true effects that a larger sample would detect. The null interaction result should therefore be interpreted cautiously: the study cannot rule out a small moderating effect of transparency that was undetectable at this sample size.

Because all participants rated both standard and VS images in a single session, the information

provided for VS images in Conditions 3 and 4 was visible throughout the session. This may have influenced ratings of standard images as well, inflating the condition main effect and making it difficult to attribute Condition 3's higher ratings to informed perception of VS images specifically. A between-subjects image type manipulation — where participants see either only standard or only VS images — would eliminate this confound.

Participants were recruited primarily through SurveySwap.io and personal networks, yielding a young, moderately AI-literate sample (86% aged 18–34). This demographic may be less susceptible to the human bias against AI-generated art documented in other populations [CMSW18], potentially underestimating the effect that transparency could have on less AI-familiar raters.

Only ten image pairs were used. The prompts were deliberately constrained to brief, generic phrases to isolate the VS manipulation, but this limits the ecological validity of the stimuli. The results may not generalise to more complex or domain-specific image generation tasks.

7 Conclusions and Further Research

This thesis investigated whether revealing the verbalized sampling prompting method causes VS-generated images to receive higher creativity ratings than standard-prompted images. The results provide partial support for this claim.

VS images were rated significantly higher than standard images on Creativity, Novelty, and Surprise across all transparency conditions, with large effect sizes ($\eta_p^2 > 0.54$). This suggests that verbalized sampling produces images that humans perceive as more creative, novel, and surprising — even when no information about the generation method is provided. The images themselves appear to carry the creativity signal.

However, transparency about the prompting method did not selectively amplify these ratings. No significant interaction between image type and transparency condition was found for any composite variable, and Condition 3 (full method explanation) did not outperform Condition 2 (elaborated prompt only). The modest condition main effect that appeared for aesthetic and functional value did not survive Bonferroni correction and likely reflects a session-wide priming effect rather than targeted improvement in VS image perception.

The honest answer to the research question is therefore: *VS images receive higher creativity ratings than standard images, but this advantage is driven by the images themselves rather than by revealing the generation method.* Transparency raises ratings broadly across a session, but does not specifically amplify the creative premium of VS images.

7.1 Future Work

Several extensions merit consideration.

Larger and more balanced samples. A replication with $n \geq 30$ per condition would provide adequate power to detect the modest condition effects suggested by the current data ($p \approx .06$–$.09$) and more precisely characterise the interaction.

Between-subjects image type manipulation. Separating participants into standard-only and VS-only groups would eliminate within-session contamination and allow a cleaner test of whether transparency specifically enhances VS image perception.

Extended transparency conditions. Testing a condition where participants are told images are AI-generated without any additional information would more cleanly separate source attribution effects from VS-specific effects [CMSW18].

Expert vs. lay raters. Moffat and Kelly [MK06] and Pasquier et al. [PBT+16] found that musical expertise modulates bias against algorithmic creativity. A parallel analysis contrasting trained visual artists with the general public is warranted.

Other creative domains. VS has been validated on text generation [ZYC+25]. Applying it to music generation or interactive story generation would test whether the perceptual creativity premium generalises beyond images.

Longitudinal effects. Whether familiarity with VS over time increases or decreases perceived creativity is an open question that longitudinal or repeated-exposure designs could address.

8 Appendix: image gallery

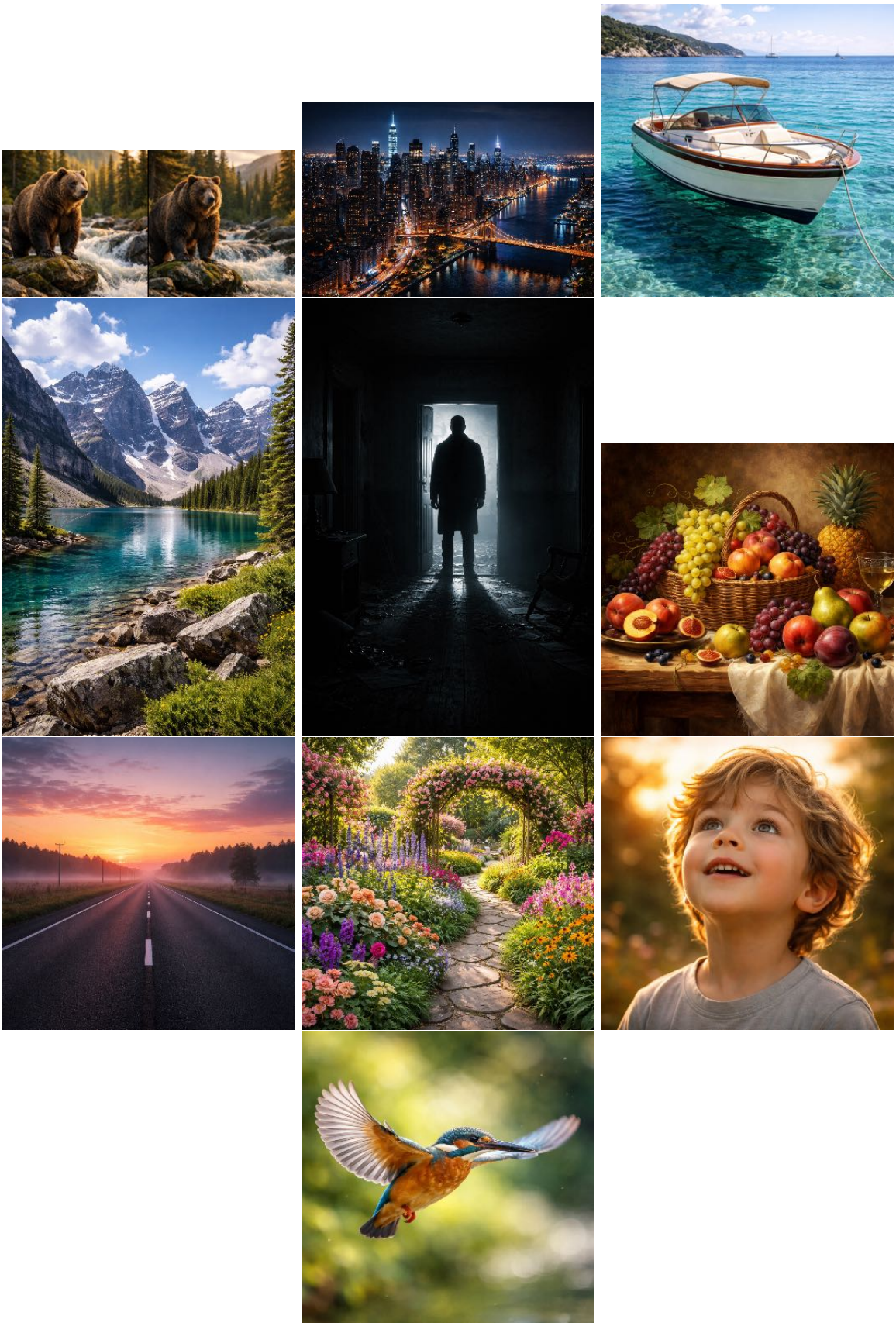

Figure 5: The ten standard images, generated directly from the input prompts (left to right, top row: bear, city, boat, mountain, doorway; bottom row: fruit, road, garden, child, bird).

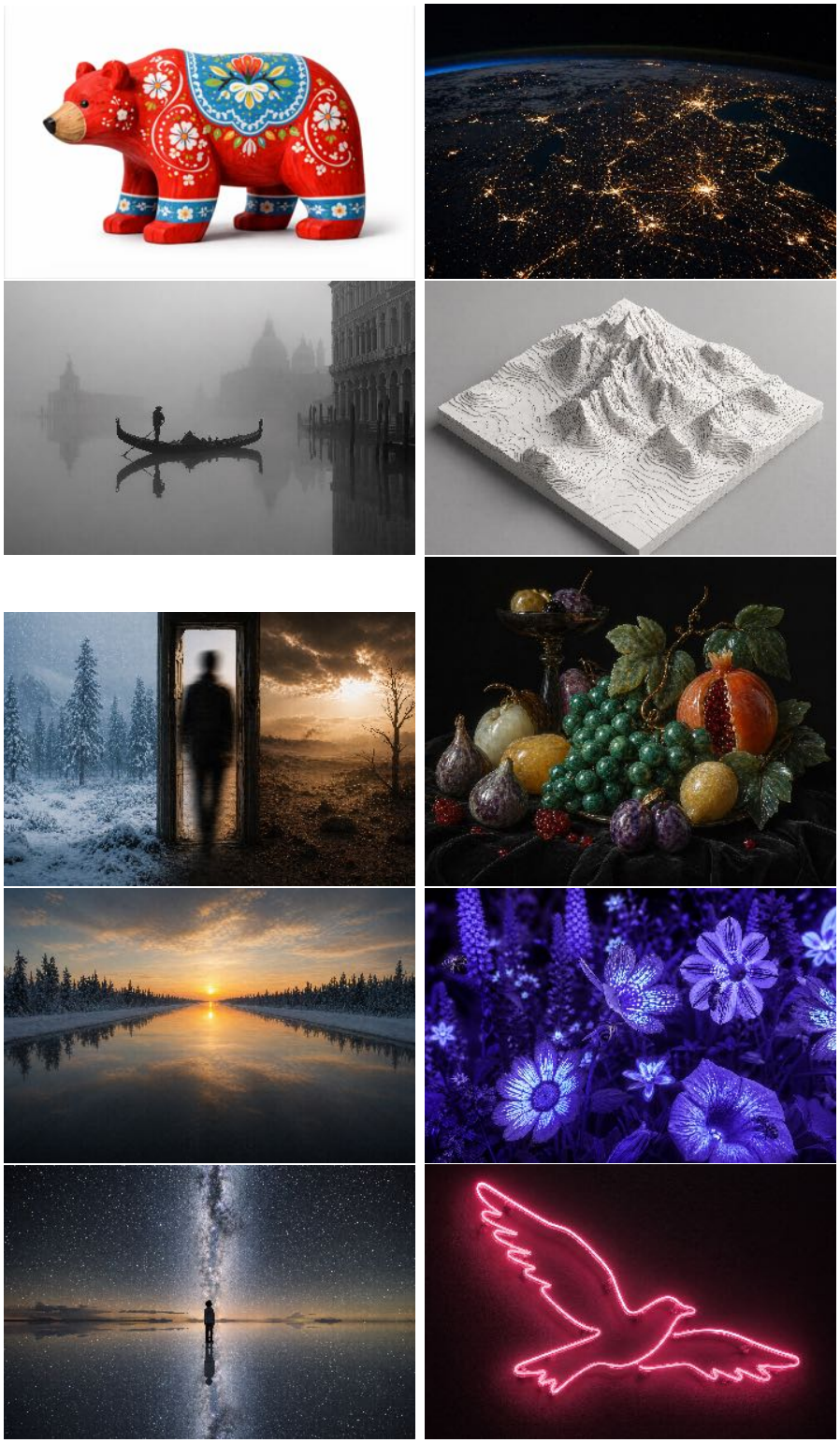

Figure 6: The ten VS images, each generated from the elaborated prompt selected by the VS pipeline (same order as Figure 5).

References

[ALM22] Ángela Agudo, Xiaolei Liu, and Patrick Magnusson. Now everybody can be an artist! bias against algorithmic art. *Computers in Human Behavior*, 2022. Reports differential evaluation of identical artworks attributed to human vs. AI creators.

[Bod04] Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

[CMSW18] Rebecca Chamberlain, Caitlin Mullin, Bram Scheerlinck, and Johan Wagemans. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2):177–192, 2018.

[CW12] Simon Colton and Geraint A. Wiggins. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pages 21–26, 2012.

[EHT23] Ziv Epstein, Aaron Hertzmann, and The Investigators of Human Creativity. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

[Gär00] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[GK19] Vlad P. Glaveanu and James C. Kaufman. The creativity matrix: Spotlights and blindspots in our understanding of the phenomenon. *The Journal of Creative Behavior*, 53(4):478–490, 2019.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

[Jor16] Anna Jordanous. Four pppperspectives on computational creativity in theory and in practice. *Connection Science*, 28(2):194–216, 2016.

[KB09] James C. Kaufman and Ronald A. Beghetto. Beyond big and little: The four c model of creativity. *Review of General Psychology*, 13(1):1–12, 2009.

[MK06] David C. Moffat and Mark Kelly. An investigation into people's bias against computational creativity in music composition. In *Proceedings of the Third Joint Workshop on Computational Creativity (ECAI 2006)*, Riva del Garda, Italy, 2006.

[MV15] Michael D. Mumford and Dan Ventura. The man behind the curtain: Overcoming skepticism about creative computing. *Journal of Creative Behavior*, 2015.

[MVD20] Maria Massi, Alessandro Vecchi, and Federica Donzelli. Innovating through storytelling: The role of process description in consumer evaluation of creative products. *Creativity and Innovation Management*, 2020. Finds that descriptions emphasising effort and deliberate work improve evaluations of creative products.

[OMD09] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[OWJ+22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022.

[PBT+16] Philippe Pasquier, Adam Burnett, Nicolas Gonzalez Thomas, James B. Maxwell, Arne Eigenfeldt, and Tom Loughin. Investigating listener bias against musical metacreativity. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 42–51, 2016.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[Rho61] Mel Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[TTT20] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[Var19] Lav R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1), 2019. Discusses the relationship between conceptual spaces and learned embedding spaces.

[WFGH11] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*, pages 143–146, New York, 2011. ACM Press.

[ZYC+25] Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025.