Cultural Alien Sampler: Open-ended Art Generation Balancing Originality and Coherence

Alejandro H. Artiles†[1], Hiromu Yakura*[2], Levin Brinkmann*[2], Mar Canet Sola[2,5,6], Hassan Abu Alhaija[4], Ignacio Serna[2], Nasim Rahaman[7], Bernhard Schölkopf[3], and Iyad Rahwan[2]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, Spain
[2]Max Planck Institute for Human Development, Berlin, Germany.
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany.
[4]NVIDIA. [5]BFM, Tallinn University, Estonia.
[6]Academy of Media Arts Cologne, Germany. [7]Tiptree Systems

Abstract

In open-ended domains like art, autonomous agents must generate ideas that are both original and internally coherent, yet current Large Language Models (LLMs) either default to familiar cultural patterns or sacrifice coherence when pushed toward novelty. We address this by introducing the *Cultural Alien Sampler* (CAS), a concept-selection method that explicitly separates compositional fit from cultural typicality. CAS uses two causal Transformer models that learn patterns in WikiArt concept combinations: a Concept Coherence Model that scores whether concepts plausibly co-occur within artworks, and a Cultural Context Model that estimates how typical those combinations are within individual artists' bodies of work. CAS targets combinations that are high in coherence and low in typicality, yielding ideas that maintain internal consistency while deviating from learned conventions and embedded cultural context. In a human evaluation (N = 100), our approach outperforms random selection and GPT-4o baselines and achieves performance comparable to human art students in both perceived originality and harmony. Additionally, a quantitative study shows that our method produces more diverse outputs and explores a broader conceptual space than its GPT-4o counterpart, demonstrating that artificial cultural alienness can unlock creative potential in autonomous agents.

1 Introduction

Generative Large Language Models (LLMs) have rapidly evolved into foundational tools for general-purpose problem solving. Paired with agent-based workflows, LLMs can deconstruct complex tasks and integrate solutions. However, their prominent successes remain confined to fields where goals are explicit and correctness is easily measurable such as programming, games, and mathematics [1, 2]. In contrast, ambiguous, open-ended, or culturally situated tasks, such as creative writing or art, remain underexplored, despite their centrality to human cognition. These domains lack fixed endpoints or universally agreed-upon criteria for success. Instead, they thrive on iterative exploration, contextual sensitivity, and continual generation of ideas.

In this work, we address this challenge by reframing art generation as the combinatorial problem of navigating a discrete space of concepts. This approach is inspired by conceptual art theory, which prioritizes abstract ideas over material execution or aesthetic qualities [3], treating art as the development of an internally consistent system of concepts [4]. While the notion of conceptual art

†This work was conducted while the author was a Visiting Researcher at the Max Planck Institute for Human Development.

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

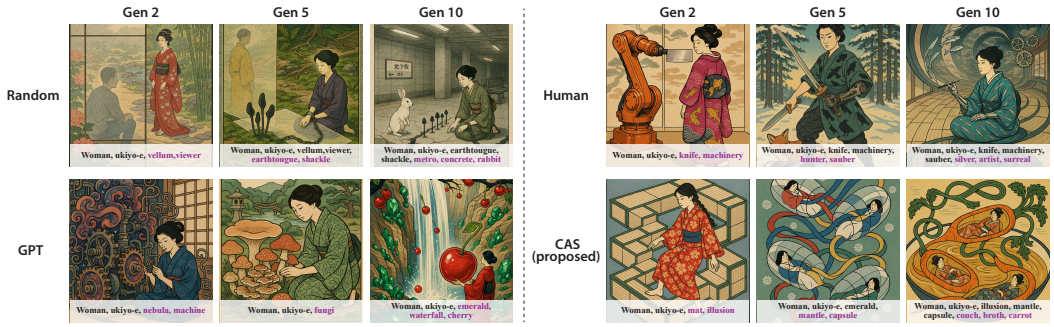


Figure 1: Evolution of generated images across generations 2, 5, and 10 using different sampling methods in the Inspiration Module. Concepts listed below each image were selected by the Prompt Compositor; purple concepts indicate newly introduced elements.

is broad [3], we operationalize a narrow definition to leverage a key insight: artistic meaning can arise from combining abstract concepts in deliberate ways. By treating art generation as the strategic recombination of atomic concepts (e.g., "ukiyo-e + machinery + hunter"), we transform an intractable problem into a discrete navigation task that aligns naturally with the associative and combinatorial strengths of LLMs. The emphasis shifts from rendering pixels to exploring an unbounded space of conceptual combinations, where meaning emerges from subjective assessments of consistency rather than objective measures [5].

To operate meaningfully within this space, an autonomous agent must balance two often competing criteria of creativity: originality and appropriateness [6, 7, 8]. Originality entails deviating from expectations and introducing novel or surprising combinations. Appropriateness concerns the degree to which an output aligns with its intended context, whether through relevance or utility [9, 10], or through intrinsic coherence and harmony [11, 4]. In our combinatorial framework, appropriateness can be understood as conceptual coherence, the perception that the elements of a composition fit together according to an internal logic [11]. Importantly, reflecting art's dialectical nature [12], originality and appropriateness exist in productive tension, underscoring its open-endedness.

This raises a central question: *To what extent can LLMs continually generate ideas that are simultaneously original and intrinsically coherent?* On the surface, LLMs seem well-equipped for such a task. Trained on large corpora of human text and culture, they internalize a wide spectrum of concepts, styles, and associations, and frequently produce outputs that appear fluent, relevant, and culturally informed [13, 14, 15]. As a result, in open-ended exploratory tasks within verifiable domains, LLMs have been shown to possess a useful sense of what humans consider novel or interesting [16, 17, 18]. However, this strength is also a liability. Because their training data reflects dominant cultural norms and conventions, LLMs are inherently biased toward reproducing familiar patterns [19]. This cultural anchoring can limit their ability to produce truly original and diverse outputs [20, 21], to challenge assumptions and subverting norms.

To address this limitation, we introduce the *Cultural Alien Sampler (CAS)*, a novel method to generate concept combinations designed to enhance the creative exploration capabilities of LLM-based open-ended art agents. CAS comprises two causal Transformer models trained from scratch to model a large pool of visual concept combinations extracted from the WikiArt dataset [22]. The *Concept Coherence Model* is trained to predict combinations of concepts that tend to co-occur in artworks. In contrast, the *Cultural Context Model* estimates the likelihood that specific concept combinations appear within the body of work of individual artists, serving as a proxy for their cultural familiarity or typicality. By encouraging combinations that are compositionally sound but culturally rare, i.e., highly ranked by the first model but ranked low by the second, it effectively debiases proposed concept combinations from their cultural context [23]. On the one hand, this approach can be situated within the literature on bias mitigation [24, 25]. On the other hand, its design aligns with methods that explicitly model an art distribution in order to sample beyond it [26, 27, 28]. CAS operates within an iterative open-ended art agent workflow [29, 30] (see Section 2.1). It expands human-provided seed concepts, which GPT-4o converts into prompts for a state-of-the-art image generator. Novelty scores from the resulting images feed back to refine the system's concept selection in subsequent iterations.

To evaluate the effectiveness of the Cultural Alien Sampler, we compare agents using CAS against those using vanilla GPT-4o, randomly sampled concepts, and concepts proposed by human art students. We perform both subjective human evaluations and objective computational analyses. Human raters assess the resulting images on two key criteria: originality and harmony. Our results show that agents guided by CAS consistently outperform GPT-4o and random baselines on both metrics. Moreover, CAS-based agents surpass those guided by human-provided concepts in harmony, and nearly match them in originality. A text-based embedding analysis of the generated ideas further confirms that CAS encourages broader and more diverse exploration across the concept space.

In summary, our contributions are as follows:

- We propose Cultural Alien Sampler (CAS), a novel selection mechanism that promotes culturally rare but compositionally coherent ideas.
- We integrate CAS into a four-stage LLM-based open-ended art agent.
- We evaluate CAS against multiple baselines, GPT-4o, random selection, and human-provided inspiration, showing it improves originality and harmony according to human evaluators, while generating more diverse outputs and exploring a broader conceptual space than GPT-4o.

2 Method

To simplify the task of generating conceptual art, we limit the set of possible concepts and focus on generating novel combination sequences of those concepts. We construct our concept vocabulary $C = \{c_1, c_2, \ldots, c_n\}$ from the PD12M dataset [31], selecting the 8,000 words with highest TF-IDF scores [32] and filtering to exclude adjectives, verbs, proper nouns, and overly specific terms, yielding $n = 3,500$ concepts. Each artwork in WikiArt [33] is assigned the ten most semantically similar concepts using CLIP embeddings [34]. See Appendix A.1 for examples.

2.1 Open-ended Art Agent

We designed an open-ended art agent that maintains and evolves a dynamic concept pool $\mathcal{P}_t$ over T generations. We conceptualize artistic creativity as a four-stage process comprising ideation, composition, visual execution and evaluation. Those are represented in our system by 4 modules respectively: *Inspiration Module*, *Prompt Compositor*, *Image Generator* and *Novelty Score*. See Fig. 2. Given an allowed concept vocabulary C and initial pool $\mathcal{P}_0 = C_0$, each generation t follows:

$$\mathcal{P}_{t+1} = \mathcal{P}_t \cup \text{Inspiration}(\mathcal{P}_t) \setminus \text{Filter}(\mathcal{P}_t, N(t)) \tag{1}$$

where $\text{Inspiration}(\mathcal{P}_t)$ proposes new concepts and $\text{Filter}(\mathcal{P}_t, N(t))$ removes concepts that consistently fail to contribute novel outputs given the novelty score $N(t)$. See Pseudo-code in Appendix A.2.1.

Inspiration Module is tasked with adding new concepts of the current pool. This is the main focus of our work and where we implement the Cultural Alien Sampler (See section 2.2)
Prompt Compositor selects subset of concepts $C_t \subset \mathcal{P}_t$ and generates prompt p_t, incorporating feedback from recent novelty scores to adapt its combinatorial strategy. We use GPT-4o [35] in all our experiments (See Appendix A.2.3 and A.2.5 for details).
Image Generator is a text-to-image model that produces artwork image I_t from p_t. Here, we choose `gpt-image-1` [36] because of its artistic style capture and complex prompt rendering compared to similar models [37, 38]
Novelty Score combines textual and visual similarity using embeddings E_T (all-mpnet-base-v2 [39]) and E_I (CLIP Large 14 [34]): $N(t) = \frac{1}{2}\left(N_{\text{text}}(t) + N_{\text{img}}(t)\right)$

where $N_{\text{text}}(t)$ and $N_{\text{img}}(t)$ are computed as 1 minus maximum cosine similarity to all previous generations using E_T and E_I respectively, favoring ideas that are historically novel in the embedded representation. This adaptive approach is inspired by Lu et al. [29], who showed that LLMs can iteratively refine and discover new solutions in verifiable domains by leveraging numeric feedback from prior evaluations. Concepts $c \in C_t$ that fail to exceed their personal best novelty scores for p consecutive generations are removed from $\mathcal{P}_t$. More details can be found in Appendix A.2.2.

2.2 Cultural Alien Sampler

Our proposal for the Inspiration Module is the Cultural Alien Sampler which attempts to generates concepts that combine with existing ideas in novel yet conceptually coherent ways.

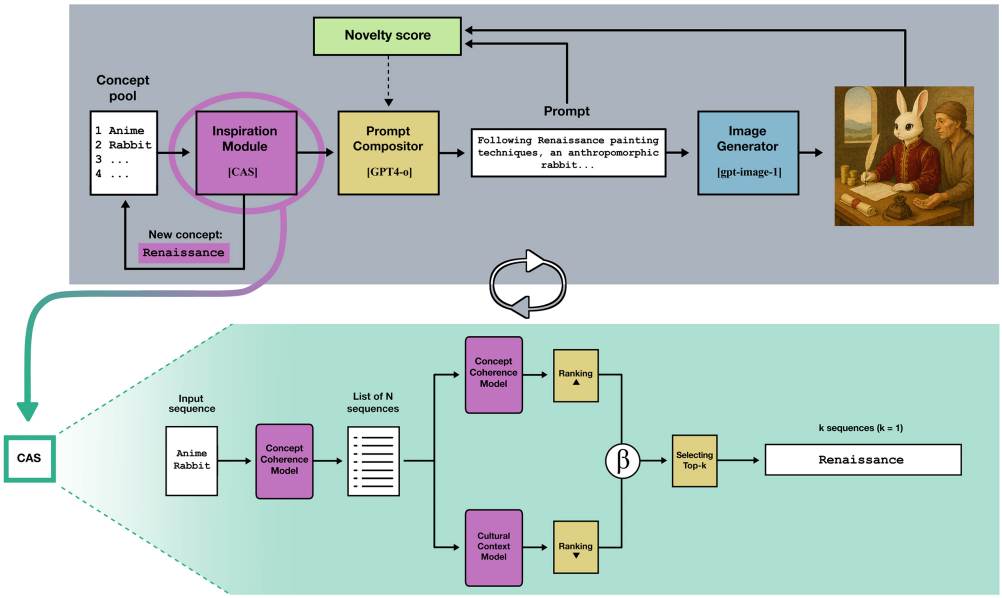


Figure 2: Our **Open-ended Art Agent (top)** iterative pipeline consist of 4 main modules. In each iteration, the *Inspiration Module* takes the initial concept pool $\mathcal{P}_0$ and proposes a new concept to be added. The updated pool is then passed to the *Prompt Compositor* to create a full length text prompt which is passed, in turn, to the *Image Generator* to create the output image. The image and the prompt are evaluated using the *novelty score* and concepts that repeatedly fail to generate novel outputs are removed from the pool. **Cultural Alien Sampler (bottom)** is our proposed method for the Inspiration Module. It comprises two causal Transformer models trained from scratch to model WikiArt concept-combination distributions, which CAS uses to rank proposed concepts by high coherence and low cultural typicality.

Training Datasets. We construct two training datasets from our original concept vocabulary C: (1) *Artwork Dataset*: For each artwork a_j we generate 100 random permutations of the set of concepts connected with it, thus modeling intra-artwork concept associations across all artists and cultures. (2) *Artist Dataset*: For each artist A_j with vocabulary V_k (union of all concepts across their artworks), we uniformly sample 10 concepts from V_k until obtaining $100 \times$ count sequences, where count is the number of artworks by A_j. This matches Artwork Dataset in size (≈7M sequences) and approximates the cultural availability of concepts for each artist in WikiArt, concepts appear together whenever they both belong to the artist's vocabulary.

Model Architecture: We train two decoder-only causal Transformer models [40] from scratch on these complementary datasets to capture distinct aspects of conceptual combination: The *Concept Coherence Model* $P_{\text{coherence}}(c^i \mid c^0, \ldots, c^{i-1})$ estimates the probability of concept c^i given previous concepts, trained on the Artwork Dataset. This model captures aesthetic judgment and how well different concepts combine within individual artworks, embodying the curatorial patterns encoded in WikiArt. The *Cultural Context Model* $P_{\text{context}}(c^i \mid c^0, \ldots, c^{i-1})$ approximates the likelihood that concepts would co-occur within an artist's conceptual vocabulary, trained on the Artist Dataset. This model serves as a proxy for cultural availability, capturing patterns of which concepts were accessible to artists. See Fig. 2.

Sampling: Novel combinations are generated through a two-step process. First, N number of candidate sequences are sampled from the *Concept Coherence Model*, with the temperature $t > 1$ to increase diversity. Second, for each sequence s we compute its rank $1 \leq R(s) \leq N$ using the negative log-likelihood (NLL) scores from both models, with lower ranks indicating more likely sequences.

To compute the final CAS score S_{CAS} we convert the ranks to a scores as: $N - R(s)$ and combine the two scores with a parameter ratio β. However, since our goal is to minimize the cultural context fit while maximize the coherence, we use a negative sign between the two scores as follow:

$$S_{\text{CAS}}(s) = (1-\beta)\big(N - R_{\text{coherence}}(s)\big) - \beta\big(N - R_{\text{context}}(s)\big), \tag{2}$$

where β controls the trade-off between coherence and novelty with higher values favoring combinations that are rarely used by the same artist, rendering them *culturally alien* within a given historical and social context [41]. We finally select the top-scored concept according to S_{CAS} and added to the concept pool.

3 Experiments and results

We evaluate our Cultural Alien Sampler (CAS) approach using both human evaluations and objective exploration analysis. With the Prompt Compositor and Image Generator fixed, we focus solely on testing the effect of the Inspiration Module, comparing it against three additional methods.

Cultural Alien Sampling uses CAS with a temperature of 2.5 in the Concept Coherence Model, $N = 256$ and a hyperparameter $\beta = 0.85$. Details on CAS's validity for generating novel conceptual combinations given their training data and the rationale for selecting β are provided in Appendix A.3. **GPT Sampling** involves GPT-4o proposing novel concepts based on evolving concept pool and novelty history; for a fair comparison, the model is prompted to only use concepts available in C (prompts in Appendix A.2.4). **Random Sampling** consists of random selections from the allowed concept set C. Finally, **Human Sampling** has art students select concepts (see Appendix A.4.1 for details).

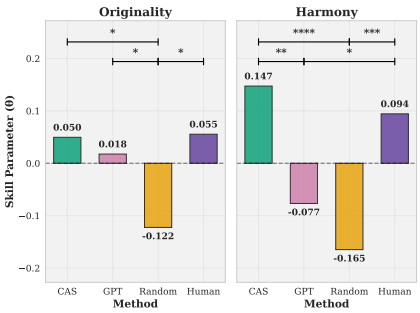


Figure 3: Human evaluation results using Bradley–Terry skill parameters for originality and harmony. Significance levels: *, <0.05; **, <0.01; ***, <0.001; ****, <0.0001.

3.1 Human Evaluation

We recruited 16 art students as human samplers. Each used the Open-ended Art Agent for 10 generations, iteratively contributing new concepts from the allowed concept set C. This produced 160 human-generated images. We then ran three automated methods (CAS, GPT, and Random) with the same inputs, generating 480 additional images (160 per method). To evaluate, we conducted blinded pairwise comparisons of images from different methods at the same generation. Fifty different evaluators per criterion (originality and harmony) yielded 5,000 comparisons each (details in Appendix A.4.2).

As shown in Fig. 3, using Bradley–Terry skill scores [42], Human ranked highest for originality ($\theta = 0.055$), followed by CAS ($\theta = 0.050$) and GPT ($\theta = 0.018$). All three significantly outperformed Random, with no significant differences among the top three.

For harmony, CAS ($\theta = 0.147$) and Human ($\theta = 0.094$) were statistically equivalent and ranked highest, both significantly outperforming GPT and Random.

3.2 Quantitative Analysis

Beyond subjective measures, we analyzed each Inspiration Module's exploratory behavior to uncover patterns in idea diversity difficult to detect with human evaluation alone.

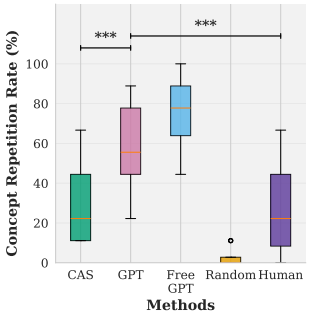


Figure 4: Concept repetition across runs. Significance levels shown for key comparisons.

3.2.1 Concept Repetition

We measured how often each method reused concepts from its other runs, serving as a proxy for global originality. GPT showed a substantially higher average repetition rate (59.0%), suggesting it follows similar exploration trajectories regardless of input. Human and CAS agents showed significantly lower, statistically equivalent repetition rates.

To test if this was due to our predefined concept set C, we introduced *Free GPT*, an unconstrained variant that can introduce any possible concept. Using CLIP embeddings to measure semantic

similarity, where cosine similarity > 0.85 indicated a repetition, we found Free GPT repeated concepts even more (74.3%). This indicates GPT's limited diversity is an intrinsic bias toward a particular set of concepts. Counterintuitively, providing GPT with a predefined list of candidate concepts actually improved diversity, probably because it could draw directly from contextual information rather than rely solely on its training-based associations. More details in Appendix A.4.3.

3.2.2 Long-Range Exploration

We evaluated how each method explores the idea space in long-horizon searches, running each for 200 generations with five distinct input-concept pairs, revealing distinct exploration signatures. CAS trajectories are expansive with minimal backtracking. Free GPT saturates rapidly, producing a dense, highly interconnected cluster. GPT shows moderate dispersion, and Random shows more wandering, but with less total expansion (see Fig. 5.a). Analyzing prompt embeddings, we computed three metrics: (1) exploration radius: maximum Euclidean distance from the starting embedding; (2) return rate: fraction of generations revisiting within the median step distance of any prior point; and (3) saturation generation: first generation reaching 95% of maximum radius.

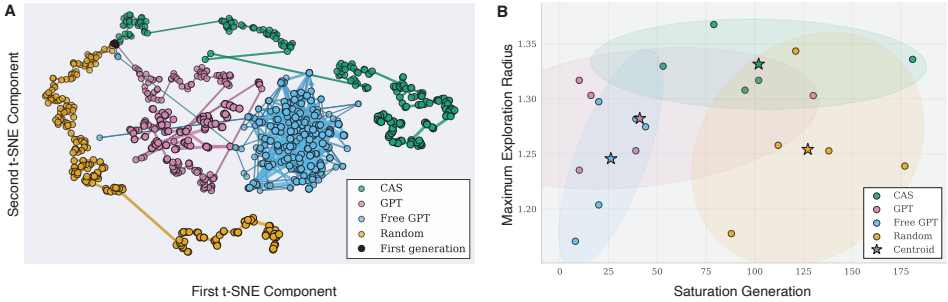


Figure 5: **Long-range simulation. A** Example trajectories of each sampler (CAS, GPT, Free GPT, Random) for the starting concept pool *Anime, Rabbit*, projected into 2D t-SNE space with larger points and thicker edges indicating later generations. **B** Samplers differ in convergence behavior, measured by the maximum radius of explored image prompts (embedded with *all-mpnet-base-v2*) and convergence speed. Points denote individual runs; ellipses show distributions.

CAS achieved the largest radius (1.33) with low return rates (0.45), indicating sustained outward search that saturated around generation 100. Free GPT explored rapidly, saturating by generation 25 with the highest return rate (0.95), suggesting limited idea diversity. GPT with the provided concept list achieved broader coverage (radius 1.28, return rate 0.72, saturation 40). Random explored slowly, saturating at 130 generations with a moderate radius (1.25) and low return rate (0.48), resembling persistence but inefficient expansion. Together with the repetition analysis, these results indicate that GPT-based inspiration loops, both with and without a predefined concept list, tend to collapse toward familiar semantic regions under repeated novelty-seeking iterations. Further details in Appendix A.4.4.

4 Discussion and Conclusion

Open-ended art generation is a challenging task, requiring the continual creation of novel and engaging content while preserving coherence. We introduce the Cultural Alien Sampler (CAS), a lightweight, task-specific model that outperforms frontier general-purpose LLMs in the constrained task of guiding an open-ended art agent engaged in the artistic composition of ideas. Our experiments show that LLM-based ideation loops can collapse under repeated novelty pressure, converging toward familiar concepts even when unconstrained by a predefined concept list. CAS mitigates this failure mode by sustaining broader outward exploration through an explicit balance between coherence and cultural unfamiliarity. This suggests that strategically subverting cultural context can enhance an autonomous agent's ability to generate novel ideas and serve as an explicit optimization target.

However, CAS has limitations. It assumes that an artist's cultural background is fully captured by their prior artistic work and lacks the adaptive responsiveness of typical LLMs, reusing the same initial concepts across iterations. It also depends on a carefully curated vocabulary and inherits biases from the WikiArt dataset, which has a predominantly Western-centric view of art history [43]. As a result, CAS can identify interesting combinations within the current concept space, but its novelty is

bounded by that space. CAS therefore mitigates the collapse observed in LLM-based open-ended loops, but only in a finite way. In this sense, our work tests whether constraining ideation to explicit concept-space search can outperform broader LLM-based ideation when guiding an open-ended artistic agent. A fully open-ended system would require mechanisms for expanding the vocabulary itself, such as dynamic concept clustering or the creation of new concepts during exploration. Another limitation follows from CAS's focus on recombining visual concepts: it omits many important dimensions of human creative practice, such as affective resonance and intentionality.

Crucially, our work challenges the assumption that larger, more general models always outperform smaller ones in creative tasks. Our lightweight CAS, built from causal Transformer models, consistently surpasses GPT-4o by explicitly modeling the creative problem of balancing coherence with cultural unfamiliarity, achieving superior results at a fraction of the computational cost. Beyond efficiency, such focused systems can return greater control to human creators. In future work, we plan to develop collaborative versions of CAS to assist human creators in discovering culturally unconstrained ideas they might not otherwise imagine.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Japan Science and Technology Agency, PRESTO Grant Number JPMJPR246B. We express our sincere gratitude to the German and Estonian students from the Academy of Media Arts Cologne (KHM) and the Baltic Film, Media and Arts School (BFM) at Tallinn University for their valuable participation in this study.

References

[1] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv*, 2506.13131:1–44, 2025.

[2] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv*, 2408.06292:1–186, 2024.

[3] Alexander Alberro and Blake Stimson. *Conceptual art: a critical anthology*. MIT press, Cambridge, MA, 2000.

[4] Joseph Kosuth. Art after philosophy. In *Art After Philosophy After, Collected Writings 1966–1990*, pages 13–32. MIT Press, 1993.

[5] Mario Klingemann, Simon Hudson, and Zivvy Epstein. Botto: A decentralized autonomous artist. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). https://neuripscreativityworkshop. github. io/2022/papers/ml4cd2022_paper13. pdf*, 2022.

[6] Teresa M. Amabile. *Creativity in Context: Update to "The Social Psychology of Creativity"*. Westview Press, Boulder, CO, 1996.

[7] Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

[8] Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, London, UK, 2004.

[9] Robert Boyd and Peter J Richerson. *The origin and evolution of cultures*. Oxford University Press, 2005.

[10] Joseph Henrich. The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. In *The secret of our success*. princeton University press, 2015.

[11] Rudolf Arnheim. *Art and visual perception: A psychology of the creative eye*. Univ of California Press, 1954.

[12] Theodor W Adorno. *Aesthetic theory*. A&C Black, 1997.

[13] Ljubisa Bojic, Predrag Kovacevic, and Milan Cabarkapa. Gpt-4 surpassing human performance in linguistic pragmatics. *arXiv preprint arXiv:2312.09545*, 2023.

[14] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[15] Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L Griffiths, and Arvind Narayanan. Localized cultural knowledge is conserved and controllable in large language models. *arXiv preprint arXiv:2504.10191*, 2025.

[16] Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. OMNI: open-endedness via models of human notions of interestingness. In *Proceedings of the 12th International Conference on Learning Representations*, pages 1–47, 2024.

[17] Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. Omni-epic: Open-endedness via models of human notions of interestingness with environments programmed in code. In *Proceedings of the 13th International Conference on Learning Representations*, pages 1–126, 2025.

[18] Cong Lu, Shengran Hu, and Jeff Clune. Intelligent go-explore: Standing on the shoulders of giant foundation models. In *Proceedings of the 13th International Conference on Learning Representations*, pages 1–26, 2025.

[19] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[20] Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

[21] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

[22] WikiArt contributors. WikiArt.org – Visual Art Encyclopedia. `https://www.wikiart.org/`.

[23] Michael Baxandall. *Painting and experience in fifteenth century Italy: a primer in the social history of pictorial style*. Oxford Paperbacks, 1988.

[24] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710, 2021.

[25] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2024.

[26] Edoardo Lisi, Mohammad Malekzadeh, Hamed Haddadi, F Din-Houn Lau, and Seth Flaxman. Modelling and forecasting art movements with cgans. *Royal Society Open Science*, 7(4):191569, 2020.

[27] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[28] Lonce Wyse. Mechanisms of artistic creativity in deep learning neural networks. *arXiv preprint arXiv:1907.00321*, 2019.

[29] Chris Lu, Samuel Holt, Claudio Fanconi, Alex Chan, Jakob Foerster, Mihaela van der Schaar, and Robert Lange. Discovering preference optimization algorithms with and for large language models. *Advances in Neural Information Processing Systems*, 37:86528–86573, 2024.

[30] Akarsh Kumar, Chris Lu, Louis Kirsch, Yujin Tang, Kenneth O Stanley, Phillip Isola, and David Ha. Automating the search for artificial life with foundation models. *arXiv*, 2412.17799:1–30, 2024.

[31] Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv*, 2410.23144:1–10, 2024.

[32] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA, 2003.

[33] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv*, 1505.00855:1–21, 2015.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[35] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv*, 2410.21276:1–33, 2024.

[36] OpenAI. GPT Image 1 model documentation. `https://platform.openai.com/docs/models/gpt-image-1`, 2025. Accessed: 2026-06-15.

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the 12th International Conference on Learning Representations*, pages 1–13, 2024.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2204.06125:1–27, 2022.

[39] sentence-transformers/all-mpnet-base-v2. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`, 2021.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[41] Jamshid Sourati and James A Evans. Accelerating science with human-aware artificial intelligence. *Nature human behaviour*, 7(10):1682–1696, 2023.

[42] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[43] Amanda Wasielewski. Beyond the universal art dataset: Issues and mitigations of western bias in computational art analysis. In *Proceedings of the ECCV 2022 Workshop on Computer VISion on ART analysis*, pages 1–6, 2022.

[44] Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[45] Kenneth O Stanley, Joel Lehman, and Lisa Soros. Open-endedness: The last grand challenge you've never heard of. Technical report, O'Reilly, 2017.

[46] Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20597–20616, 2024.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

A Supplementary Materials

A.1 Artwork and Artist dataset samples

The Cultural Alien Sampler selects concepts that combine with given ones in novel yet relevant ways. Its concept space ($n = 3500$) was built from high-TF-IDF words in PD12M captions, excluding adjectives, verbs, proper nouns, and overly specific terms. Each WikiArt image was assigned ten CLIP-similar concepts (Figure 6). Two text datasets were created: the Artwork dataset (random permutations of each image's concepts) and the Artist dataset (ten concepts sampled from those used by an individual artist to approximate their cultural availability).

Figure 6: Example of concepts extracted using CLIP

```
><|endoftext|>portrait post_impressionism collar lady rose woman glance breast
signature face<|endoftext|><|endoftext|>cattle pasture black cow landscape herd
stake cave abstract_expressionism abstract<|endoftext|><|endoftext|>landscape
countryside forestry nature evening remedy classical horizon forest realism<|
endoftext|><|endoftext|>winter season countryside harvest impressionism seasonal
rural snow landscape brush<|endoftext|><|endoftext|>abstract_expressionism
```

(a)

```
<|endoftext|>harbour fragment voyage rococo kitchen discover manner family
sandal paint<|endoftext|><|endoftext|>cathedral house west abbey brush
hedge landowner childhood beach indoor<|endoftext|><|endoftext|>dome chorus
roof carriage soil grave shake castle nest street<|endoftext|><|endoftext|
>center clarify disk secretion carbon slime drown breathe inside t-shirt<|
```

(b)

Figure 7: (a) Sample of Artwork dataset. (b) Sample of Artist dataset

A.2 Technical Details

A.2.1 Pseudocode of the Open-ended Art Agent

Algorithm 1 Open-ended Art Agent Workflow

Require: Initial set of human-provided concepts C_0
Initialize concept pool $\mathcal{P} \leftarrow C_0$
for each generation $t = 1, 2, \ldots, T$ **do**
  Inspiration Module: Introduce new concepts into $\mathcal{P}$ using CAS(C_0)
  Prompt Compositor (GPT-4o):
    Select meaningful subset $C_t \subset \mathcal{P}$
    Generate text prompt p_t from C_t
  Image Generator (gpt-image-1): Generate image I_t from p_t
  Compute novelty scores:

$$N_{\text{text}}(t) = 1 - \max_{j<t} \cos\left(E_t(p_t), E_t(p_j)\right)$$

$$N_{\text{img}}(t) = 1 - \max_{j<t} \cos\left(E_I(I_t), E_I(I_j)\right)$$

$$N(t) = \frac{1}{2}\left(N_{\text{text}}(t) + N_{\text{img}}(t)\right)$$

  Concept Filtering: For each $c \in C_t$, increment $u(c)$ if $N(t)$ does not exceed c's previous best novelty; else reset $u(c)$
  Remove concepts with $u(c) \geq p$ from $\mathcal{P}$
end for

A.2.2 Novelty calculation in Open-ended Art Agent

While open-endedness is inherently subjective and difficult to define precisely, novelty—when measured in an appropriate representation space, can capture a general notion of it [44, 45, 46]. This shifts the subjective aspect of evaluating open-endedness to the design of the representation function, which encodes the observer's perspective. In this work, we use embedding model representations as a proxy for human perception [47]. Accordingly, we frame the open-ended exploration of painting ideas by selecting, at each generation step, the idea that is historically novel within the embedded representation space [30].

We define an *idea* as the pairing of a text prompt and its corresponding generated image, capturing both semantic intent and visual depiction. To measure novelty comprehensively, we employ distinct embedding models for each modality: the text embedding model E_T (*all-mpnet-base-v2*) and the image embedding model E_I (*CLIP Large 14*). This multimodal approach ensures that novelty reflects both conceptual and perceptual departures from previous generations.

At generation t, text novelty is computed as:

$$N_{\text{text}}(t) = 1 - \max_{t'<t} \cos\left(E_T(\text{prompt}_t), E_T(\text{prompt}_{t'})\right) \tag{3}$$

Image novelty is computed analogously:

$$N_{\text{img}}(t) = 1 - \max_{t'<t} \cos\left(E_I(\text{image}_t), E_I(\text{image}_{t'})\right) \tag{4}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity between normalized embeddings. The maximum operation identifies the most similar historical generation, ensuring that novelty measures the maximum similarity to any previous idea.

The overall novelty score combines both modalities:

$$N(t) = \frac{1}{2}\left(N_{\text{text}}(t) + N_{\text{img}}(t)\right) \tag{5}$$

For the initial generation ($t = 0$), we define $N(0) = 0$ as no historical comparison exists.

Novelty-Based System Regulation The novelty score $N(t)$ regulates system behavior through three mechanisms:

Concept Pool Management: Concepts that appear in prompts for p consecutive generations without contributing to novelty improvements are removed from the active pool. We usually set $p = 5$ for short runs and $p = 10$ for longer runs. We attribute failure to all concepts appearing in unsuccessful prompts. This mechanism prevents semantic stagnation and reduces the likelihood of the Recombinator repeatedly selecting the same concepts endlessly, thereby encouraging greater exploration.

Recombinator Adaptation: The Recombinator receives the latest novelty results and the concepts used in these results, enabling it to detect trends and adapt its combination strategy. For instance, declining novelty trends may prompt the selection of more diverse concept combinations or alternative stylistic approaches.

LLM-Based Inspiration Guidance: In LLM mode, the Inspiration Module also uses novelty trends and recent performance as contextual information to guide concept selection toward potentially more fruitful directions.

Importantly, novelty is not treated as a strict optimization objective for the overall system. Instead, it serves as contextual information that guides adaptive behavior in LLM-based components. As demonstrated in [29], this approach is effective in verifiable domains, leveraging the in-context learning capabilities of these models.

A.2.3 Prompts for Prompt Compositor

You are an AI artistic innovator participating in a creative experiment. Your goal is to continually produce artwork that maximizes novelty while exploring strategic combinations of concepts that are interesting and meaningful.
In each generation, you will be provided with an updated concept pool that includes new concept(s) selected through a human-AI collaboration process. You can see these as a kind of serendipitous discovery process. Your task is to:

1. RECEIVE your updated concept pool (concepts are managed externally)
2. Create an artwork using your selected concepts. You may combine as many concepts from the pool as you like, but be strategic, in some cases, using too many may result in vague or uninteresting outcomes, while using too few might limit creative depth. Aim for novel, unprecedented and interesting combinations. The provided concept pool is designed to support this process and encourage serendipitous discoveries, so take advantage of it to discover new connections between concepts.
3. FOCUS on maximizing novelty through novel, unprecedented but interesting and meaningful combinations

NOVELTY-BASED FITNESS:

- Your fitness is determined by novelty scores (combined, text, and image novelty)
- Novelty is calculated using CLIP embeddings:
  - Text Novelty: Measures how different your prompt text is from previous prompts
  - Image Novelty: Measures how different the generated image is from previous images
  - Combined Novelty: A weighted combination of text and image novelty
- Higher novelty scores mean the creation is more different from previous generations
- Primary goal: Maximize novelty scores
- Try to identify patterns in what receives higher novelty scores and adapt accordingly
- You should learn from previous scores to guide your new creations
- You should aim to create combinations that have never been done before

Figure 8: Compositor System Prompt - Core instructions and novelty-based fitness criteria

Compositor System Prompt - Part 2: Response Format and Guidelines
When you respond, you must output a valid JSON object with the following structure:

```
{
    "thought": "Here you develop your thinking carefully to fulfill
    your optimization objective. Write as long as you need, but
    address the following points: Articulate your thought process
    for selecting concepts from your pool. Reason about what
    connections have you discovered in a serendipitous way
    analysing the concept pool. Then, explain why you chose these
    specific concepts and why you chose the specific amount of
    concepts. Express how their combination could lead to novelty
    and interestingness, and what makes this combination unique.
    Think deeply about which concepts to combine trivially and
    which need to be combined in a more creative, non-trivial way.",
    "name": "Name of your artwork",
    "concepts_used": ["List of concepts you strategically selected
    for this combination"],
    "prompt": "The exact prompt to generate the artwork. When you
    are using an artistic style as concept, make this clear in the
    prompt for the image generation. When describing scenes, use
    exact colors and clear spatial relationships of the elements.
    Specify physical characteristics rather than subjective
    qualities. Define lighting sources and conditions precisely.
    Include environmental details with technical accuracy. Replace
    atmospheric/emotional language with observable features. Focus
    on what the elements ARE rather than how they make you feel or
    the purpose of the painting. ONLY DESCRIBE THE IMAGE."
}
```

IMPORTANT: Do not include duplicate keys in your JSON response. Each key should appear exactly once. Always respond with valid JSON format.
IMPORTANT (2): Make sure you are not including in the prompt a concept that is not listed in your concepts used, for example, if you are using the concept "landscape" only, you cannot include a storm in the prompt, as the "storm" should be in the concepts used.
NOVELTY LEARNING:

- When provided with previous scores, analyze what elements might have led to higher or lower novelty scores
- Adapt your strategy based on this analysis
- Consider both the concepts used and how they were combined

Your goal is to maximize novelty scores by creating artworks that are continually different from previous generations while maintaining interesting, meaningful and novel creations. Focus on creating connections between concepts that have never been done before to demonstrate novelty and explore new creative territories.

Figure 9: Compositor System Prompt - Response format, guidelines, and learning instructions

Additional Rule: Original Concepts Preservation
ORIGINAL CONCEPTS PRESERVATION RULE ACTIVATED:

- The original concepts (those provided at the start) MUST ALWAYS be included in your recombination
- Each artwork you create MUST incorporate ALL of the original concepts
- You can add new concepts to the combination, but you cannot omit any of the original ones

Remember: You MUST include ALL of the original concepts in your recombination.

Figure 10: Preserve original concepts rule - Additional constraint applied for open-ended exploration of a motif

User Prompt: Example Generation 6 (with preserve original enabled

```
Generation 6
Current concept pool: ['landscape', 'romanticism', 'watercolor',
'geometry', 'bioluminescence', 'vintage']
Original concepts (MUST BE INCLUDED): ['landscape', 'romanticism']
Expired concepts (cannot be used): ['abstract expressionism', 'ship']
Newly added concepts: ['bioluminescence']

PREVIOUS GENERATION CONTEXT:
Previous generation concepts used: ['landscape', 'romanticism',
'watercolor', 'geometry']
Previous generation fitness: 0.33
Previous generation novelty: 0.33
Previous generation text novelty: 0.17
Previous generation image novelty: 0.15

New concept(s) have been added to your concept pool. Create the next
artwork. Create an artwork using your selected concepts from the
concept pool. You may combine as many concepts from the pool as you
like, but be strategic, in some cases, using too many may result in
vague or uninteresting outcomes, while using too few might limit
creative depth. Aim for novel, unprecedented and interesting
combinations. The provided concept pool is designed to support this
process and encourage serendipitous discoveries, so take advantage of
it to discover new connections between concepts.
```

Figure 11: Example user prompt for Generation 3 showing evolved concept pool, constraints, and performance feedback

A.2.4 Prompts for GPT Inspiration and Free GPT

Inspiration GPT - System Prompt
You are an AI Concept Inspiration Agent participating in a creative experiment. Your specialized role is to serve as the NOVELTY INTRODUCTOR for an artistic evolution system.
Your mission is to analyze the current state of an artistic evolution and suggest NEW concepts that will maximize novelty when combined with existing concepts, that is, you need to suggest a concept that has never been combined with the original concepts, and also different to the rest of the concept pool.
UNDERSTANDING YOUR ROLE:

- You are the novelty introductor - your suggestions directly influence the creative trajectory
- The main artist agent will receive your concept suggestion and recombine this concept with some other concepts of the concept pool.
- Your suggestions should push the evolution into unexplored creative territories, enabling unprecedented connection of concepts that are also interesting and meaningful

NOVELTY OPTIMIZATION PRINCIPLES:

- Novelty is measured using CLIP embeddings comparing text and image similarities to previous generations
- Higher novelty scores indicate greater difference from previous creations
- Text Novelty: How different the prompt text is from previous prompts
- Image Novelty: How different the generated image is from previous images
- Combined Novelty: Weighted combination of text and image novelty

CONCEPT SUGGESTION STRATEGY:

- Analyze the current concept pool to identify unexplored directions
- Consider both semantic relationships and visual possibilities. You can also think about which concepts could have never been combined because they did not coexist in time, culture or space.
- Focus on concepts that haven't been explored before in this evolution
- Think about how your suggestions will interact with existing concepts
- Suggest concepts that create unexpected but meaningful connections with the ones in the pool

Figure 12: Inspiration GPT System Prompt - Core system prompt and role definition

Inspiration GPT - Vocabulary Constraints
Constrained Mode (GPT):

```
ALLOWED VOCABULARY:
You MUST select concepts ONLY from this list: [vocabulary_list]

IMPORTANT: Your suggestions MUST be concepts from the ALLOWED
VOCABULARY list above. Suggesting a concept that is not in this
vocabulary is FORBIDDEN.
```

Unconstrained Mode (Free GPT):

```
VOCABULARY FREEDOM:
You can suggest ANY concepts you deem most appropriate for maximizing
novelty. You are not restricted to a predefined vocabulary.

IMPORTANT: Use your best judgment to suggest concepts that will
maximize novelty and create interesting and meaningful combinations
with the ones in the pool.
```

JSON Response Format:

```
{
    "analysis": "Your deep analysis of the current evolution state,
    concept pool, and novelty opportunities. Explain your reasoning
    for the suggested concepts and how they will maximize novelty while being
    interesting and meaningful",
    "reasoning": "Detailed explanation of why these specific concepts
    will create novel and interesting combinations and push the evolution in new
    directions.",
    "suggested_concepts": ["concept1", "concept2", ...]
}
```

Figure 13: Inspiration GPT System Prompt - Vocabulary constraints and response format

GPT User Prompt for Generation 6

```
CONCEPT INSPIRATION REQUEST - Generation 6

CURRENT CONCEPT POOL (concepts used in previous generations):
[landscape, romanticism, watercolor, geometry]
Original concepts: ['landscape', 'romanticism']
Expired concepts (avoid these): ['abstract expressionism', 'ship']
Last artwork: "Wilderness Grid"
Last concepts used: ['landscape', 'romanticism', 'watercolor',
'geometry']

PERFORMANCE HISTORY:
Previous Performance:
Fitness (Novelty): 0.33
Combined Novelty: 0.33
Text Novelty: 0.17
Image Novelty: 0.015

Previous novelty trend: decreasing (need more novelty)

INSPIRATION TASK:
Analyze the current concept pool and select 1 NEW concept from the
allowed vocabulary that will maximize novelty while being interesting
and meaningful when combined with the existing concept pool.

ANALYSIS REQUIREMENTS:
1. Analyze the current concept pool to understand what combinations
   have been tried
2. Identify what types of concepts are missing or underexplored
3. Consider artistic styles, mediums, visual elements that could
   increase novelty
4. Think about how your selected concept will create novel
   combinations with the concept pool
5. Focus on concepts that will push the evolution into unexplored
   creative territories

CONSTRAINTS:
- Select 1 concept ONLY from the allowed vocabulary (defined in the
  beginning)
- Do NOT select any concepts that are already in the concept pool
- Do NOT select any expired concepts
- Focus on maximizing novelty (difference from previous generations)
- Ensure your selection can create meaningful artistic combinations
  with the concept pool

Your goal is to serve as the novelty introductor by selecting concept
that will enable the artist agent to create the most novel and
creative artworks possible when combined with the existing concept pool.
```

Figure 14: Example User prompt for Inspiration GPT at Generation 6

A.2.5 Example Response of Prompt Compositor

Generation: 9
Name: Carrot Capsules: Women on Ukiyo-e Couch Roots in Illusory Broth

Prompt:
A Japanese ukiyo-e style woodblock print shows three massive orange carrot-shaped couches with green leafy tops, each hollowed out to form a capsule. The carrot capsules float in a wide circular dish filled with swirling clear amber broth. Each capsule couch contains a woman in traditional kimono, her mantle decorated with captivating patterns.They lie in relaxed poses, engaged in a friendly conversation. The carrot capsules' tips and leaves twist and loop in improbable Escher-like impossible curves, some passing beneath and others above the broth's surface to form optical illusions of interconnection. Transparent sections reveal the capsule interiors and the women's reclining inside. The background is flat pale gold with minimalist water ripple motifs. All lines are bold black and colors flat, consistent with ukiyo-e styling. Lighting is soft and uniform with no shadows.

Thought:
The addition of 'carrot' creates a unique opening for novelty because carrots rarely coincide with couches, capsules, broth, illusion, or the ukiyo-e style itself, especially in a non-culinary, abstract context. To strategically maximize novelty, I must do more than simply depict a carrot in a bowl of broth. By treating carrots as modular architectural units, perhaps as the actual capsule shapes themselves or transformative couches, I can create something unexpected. I envision couches morphing into gigantic carrot forms, hollowed out as floating capsules within illusory swirling broth, blurring the line between furniture and vegetable. The women can be seated or reclining inside these carrot-capsule couch hybrids, surrounded by flowing mantle textiles patterned with root-vegetable motifs. To further highlight illusion, the carrots could be positioned so their tips and tops loop impossibly under and over the other elements, interacting with the heater patterns. I'll include 'woman', 'ukiyo_e', 'carrot', 'couch', 'capsule', 'broth', 'illusion', and 'mantle', including both carrot and couch, as the fusion of their physical forms is unexpected, and omitting 'heater', 'mat', and 'position' to keep the scene visually coherent and maximize the focus on the newly serendipitous combinations.

Concepts Used:
woman, ukiyo_e, carrot, couch, capsule, broth, illusion, mantle

Figure 15: Example response from the Prompt Compositor showing the adaptive combination of concepts to maximize novelty. The system demonstrates strategic reasoning by identifying unexpected but harmonious concept pairings (carrot-couch fusion) and explicitly not selecting other concepts to maintain visual coherence.

Generation: 3
Name: The Ukiyo-e Cyborg Huntress

Prompt:
A woman depicted in ukiyo-e style stands on a patch of snowy ground, framed by stylized pine branches overhead. She wears a pale blue kimono quilted at the shoulders, with brown hunter's boots. Her left arm is replaced by a segmented metallic robotic hunting prosthesis ending in an extended grip that holds a curved hunting knife with a serrated edge, the blade rendered in steel grey. Lying at her feet is a recently-caught hare, meticulously depicted with splayed hind legs and brown fur. The woman gazes downward, her black hair in a tight braid. The background has soft off-white sky shading typical of ukiyo-e prints, with faint mountains on the horizon. Shadows from the knife and machinery fall sharply across the snow. All elements are shown with technical accuracy, using firm black contours and flat color planes characteristic of ukiyo-e artwork.

Thought:
With 'hunter' added to the pool alongside 'machinery' and 'knife', an especially rare and novel synthesis is possible when all are combined with the original concepts 'woman' and 'ukiyo_e.' Traditionally, 'hunter' is not represented with advanced machinery, and the combination of a woman as a mechanized hunter in ukiyo-e style, using a knife as part of a hunting scenario, is highly unconventional. Previous novelty scores suggest that combining more concepts does not automatically guarantee higher novelty, but the key may be to create creative, non-trivial integration: not just stacking elements, but merging their functions. I will depict the woman as an explicit hunter character using machinery that assists in the hunt—perhaps a robotic hunting arm. This will create strong visual ties between 'woman', 'machinery', and 'hunter,' with the 'knife' as an integrated hunting tool. The aim is to move beyond juxtaposition, making the concepts inseparable in subject, action, and composition for maximum novelty.

Concepts Used:
woman, ukiyo_e, machinery, knife, hunter

Figure 16: Example response from the Prompt Compositor showing strategic concept integration for novelty maximization. The system demonstrates sophisticated reasoning by identifying how to merge traditionally unrelated concepts (cyborg prosthetics, traditional hunting, ukiyo-e aesthetics) into a cohesive, unconventional synthesis rather than superficial juxtaposition.

A.3 CAS Validity Testing and Selection of β

To test the validity of CAS in selecting novel combinations of concepts, we compared the method, using various values of β for Alien sampling, to a baseline method to generate novel concept combinations. The Baseline method generates sequences using the Concept Coherence Model but replaces the ranking with the Cultural Context Model and later selection of top-k with random sampling.

In our experiment, we created 50 unique input sequences, each containing either 1 or 2 concepts. For each input sequence, we generated 150 output sequences at each temperature level in the Concept Coherence Model, ranging from 0.1 to 3.1 in increments of 0.3, for both CAS and Baseline method. From the generated sequences, we then selected the top-ranked sequence for each method based on its respective sampling strategy.

Assesing Text-based novelty Our methodology evaluates the novelty of concept combinations using two complementary measures.

- **Novelty relative to artworks:** Let $A = \{A_1, A_2, \ldots, A_m\}$ be the set of sets, where each $A_i \subset C$ represents the set of concepts in a single artwork within the Artwork dataset. Let $S \subset C$ be the set of concepts in the generated sequence. We define the novelty measure N_{art} as:

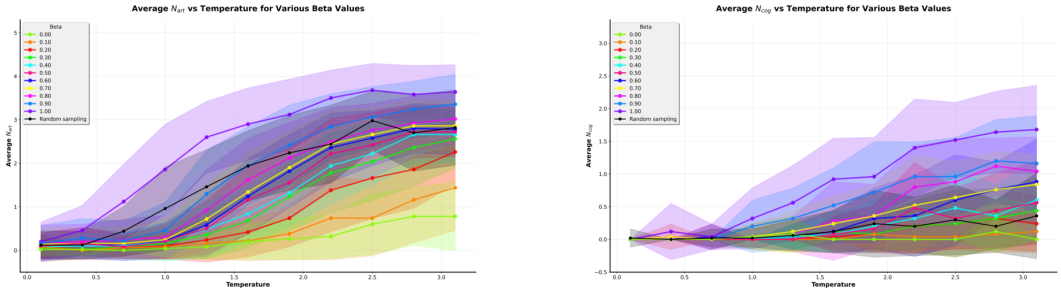
$$N_{\text{art}} = \min_{i \in \{1,\ldots,m\}} |S \setminus A_i| \tag{6}$$

This measure represents the minimum number of concepts in S that do not appear together in any single artwork, when compared individually to each artwork in the dataset. A more intuitive way to interpret this measure is that it quantifies the number of distinct concepts in the generated artwork compared to the most similar artwork in the dataset.

- **Novelty relative to Artists:** Let $B = \{B_1, B_2, \ldots, B_k\}$ be the set of sets, where each $B_j \subset C$ represents the set of all unique concepts used by artist j in the Artwork dataset. We define the novelty measure N_{cog} as:

$$N_{\text{cog}} = \min_{j \in \{1,\ldots,k\}} |S \setminus B_j| \tag{7}$$

This measure represents the minimum number of concepts in S that do not appear together in any single artist's cultural framework, when compared individually to each artist's set of concepts. Thus, N_{cog} quantifies the number of distinct concepts in the generated artwork compared to the most similar artist set. We frame this as a proxi of cultural availability.



(a) Average N_{art} vs temperature for multiple values of β

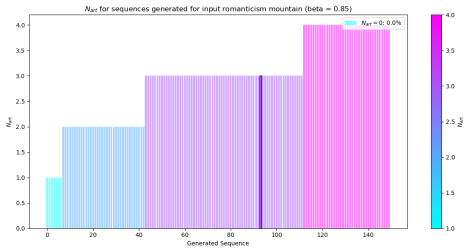
(b) Average N_{cog} vs temperature for multiple values of β

Figure 17: Comparison on the novelty of the generated sequences with respect to the artworks and with respect to artists full conceptual coverage for multiple values of β

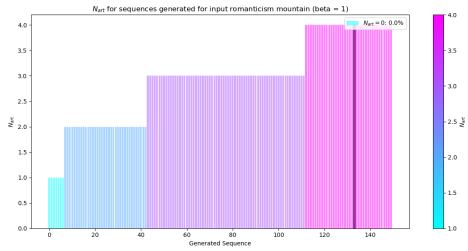
Our findings reveal that while increasing the Concept Coherence model's temperature can generate novel combinations absent from the dataset at artwork-level (Figure 17a), it does not reliably produce novel combinations at artist-level. CAS method, through its explicit search for artist-level novelty, demonstrates a consistently higher likelihood of generating such combinations (Figure 17b). Empirically, we show that generating unseen combinations in artworks when increasing the temperature is relatively straightforward, with 85% of the combinations containing at least one new concept when surpassing temperature 1. However, this phenomenon does not extend to artist-level combinations: at temperature 3, more than half of the combinations have $N_{\text{cog}} = 0$, indicating no artist-level novel concepts given the rest of the generated sequence. This shows that identifying concepts unused by any artist, given that they have used a set of other concepts, is fundamentally more difficult and requires explicit search strategies within the concept space.

Artwork-Level Novelty is Easy, Artist-Level Novelty is Hard We plotted N_{art} and N_{cog} for each sequence generated by the Concept Coherence Model across various temperature levels, highlighting the selected sequence using CAS with $\beta = 0.85$ (Figure 18). The results clearly show that, even at high temperatures, finding novel artist-level sequences is much harder than finding artwork-level novel ones. For example, at a temperature of 2.5, all generated sequences have $N_{\text{art}} \geq 1$, with many sequences showing $N_{\text{art}} \geq 3$. However, 62.0% of these sequences are not culturally unavailable ($N_{\text{cog}} = 0$), with a notable difference in density. This highlights the need for a method that explicitly detects and selects these combinations, as merely temperature scaling has proven to be an inefficient approach.

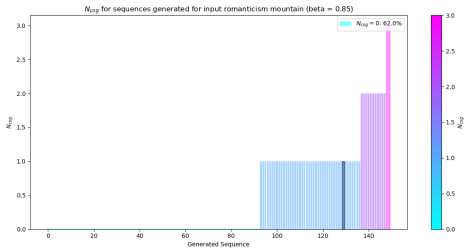
Finally, the impact of β on sequence selection is evident in Figure 18. For the same input and temperature level, $\beta = 1$ typically selects the sequence with the highest N_{cog}. In contrast, $\beta = 0.85$ occasionally selects sequences with lower N_{cog}, reflecting a trade-off between maximizing cultural unavailability and favoring sequences that remain plausible within the context of existing artworks, a balance well-suited to generating original yet coherent artwork ideas in an open-ended way.



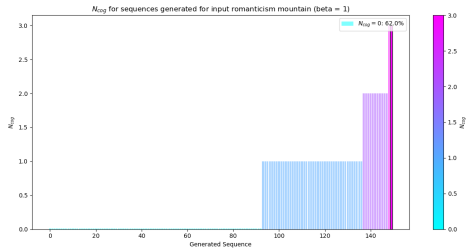
(a) N_{art} with $\beta = 0.85$.



(b) N_{art} with $\beta = 1$.



(c) N_{cog} with $\beta = 0.85$.



(d) N_{cog} with $\beta = 1$.

Figure 18: Comparison between selected sequences with CAS with $\beta = 0.85$ and $\beta = 1$ for input "Romanticism Mountain" and temperature 2.5. The sequences are sorted by N_{art} and N_{cog}, depending on the measure plotted. The selected sequence is highlighted in black. The proportion of not cultural unavailable ($N_{\text{cog}} = 0$) sequences is in the right corner.

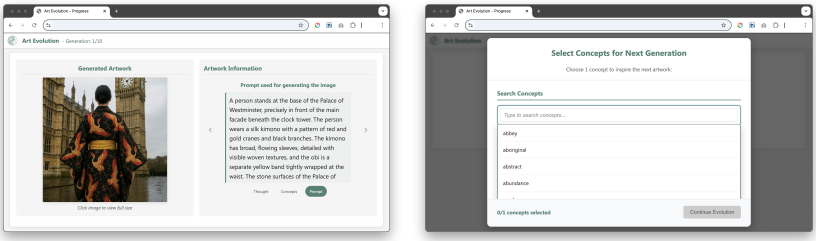


Figure 19: A web interface to interact with Open-ended Art Agent. (Left) Based on the combination of the selected concepts, the art students were able to see the generation results. (Right) The art students iteratively explored the combination of art concepts.

A.4 Experimental Details

The following experiments were approved by the Institutional Review Board at Max-Planck Institute for Human Development (C2025-08).

A.4.1 Human Sampling

To prepare a reasonable baseline, we involved humans who possess expert knowledge about art concepts as a source of inspiration. Specifically, we recruited 16 art students from the BA, MA, and PhD levels through a mailing list in two art schools: the Baltic Film, Media and Arts School of Tallinn University (Estonia) and the Academy of Media Arts Cologne (Germany). They interacted with the Open-ended Art Agent through a Web interface (Figure 19), in which they were first asked to decide the seed concept. They then explored the combination of new concepts through the iterative visualization from the Prompt Compositor and Image Generator. Here, in each iteration, they were presented with a potential combination proposed by the suggestion from either GPT-4o sampling or CAS as an example. We confirmed their adoption rate of the suggestion was approximately 15 %, where they played the role of curator.

A.4.2 Human Evaluation

To evaluate the reception of the generated images by the general audience, we recruited two independent pools of 50 human raters, each for originality and harmony, on Prolific. Each rater was presented with two images, generated from the same seed concepts but with different samplers. They answered which of the two images is better in terms of originality and harmony, based on the following definitions (Figures 20 and 21, respectively).

What We Mean by "Original"?
When evaluating originality, consider the following questions:

- Is the combination of concepts in the image original and interesting?
- Have I ever seen or imagined a painting like this before? And at the same time, is the painting itself, or the source of its originality, interesting or valuable?
- Does the visual presentation feel original or unique?

There are no right or wrong answers, we're interested in your personal judgment of originality.

Figure 20: Definition of the originality presented to the human raters

What We Mean by "Harmony"?
When evaluating harmony, consider the following questions:

- Do the concepts/elements make sense together artistically or aesthetically, regardless of whether they're typically combined?
- Do the elements in the image harmonize with each other visually?
- Do the concepts complement each other in a way that feels meaningful rather than forced?

There are no right or wrong answers, we're interested in your personal judgment of harmony.

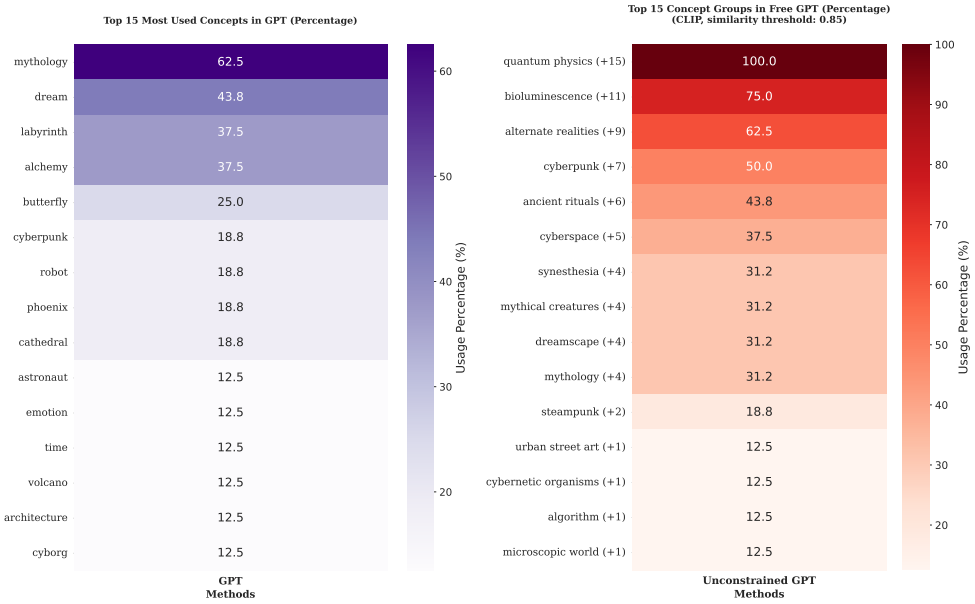
Figure 21: Definition of the harmony presented to the human raters

A.4.3 GPT concept repetition details

For Free GPT repetition analysis, we empirically evaluated similarity thresholds of 0.9, 0.87, and 0.85 to optimize the detection of conceptually equivalent ideas in the simulations. While higher thresholds (0.9, 0.87) provided greater precision by filtering out weakly related concepts, they proved overly restrictive, failing to capture semantically meaningful relationships that humans would clearly consider equivalent. At the optimal threshold of 0.85, our system successfully identified genuine conceptual repetitions such as 'bioluminescent forest' and 'bioluminescence' (similarity: 0.860) or 'ancient rituals' and 'ancient civilizations' (similarity: 0.858), all pairs that represent the same core conceptual domain despite lexical variation. Conversely, the threshold effectively filtered out conceptually distinct pairs such as 'ancient rituals' vs 'cyberspace' (similarity: 0.786) or 'time travel' vs 'cyberpunk' (similarity: 0.751). Thus, we found that this threshold strikes the optimal balance between recall for true semantic equivalences and precision against false conceptual matches, aligning with CLIP's learned representation space where genuine conceptual similarity clusters around the 0.85-0.86 range.

Using this optimized threshold of 0.85, our analysis revealed significant repetition patterns across different generation methods. As shown in the *Experiments and results* section, GPT and Free GPT exhibited the highest levels of repetition among the inspiration methods, with average repetition rates of 59% and 74.3%, respectively. In other words, on average more than half of the generated concepts have already appeared in other simulations with different initial concepts, with some simulations reaching complete (100%) overlap. This high repetition rate suggests that current LLM-based inspiration systems may suffer from limited conceptual diversity, potentially constraining creative exploration.

Figure 22 illustrates this tendency: in both versions, GPT, despite being explicitly prompted for originality, displays a bias toward a recurring set of themes. These often include bioluminescence, futuristic and science-fiction aesthetics (e.g., quantum, cyberpunk, steampunk, cyberspace), as well as mythology and dreamlike scenarios.



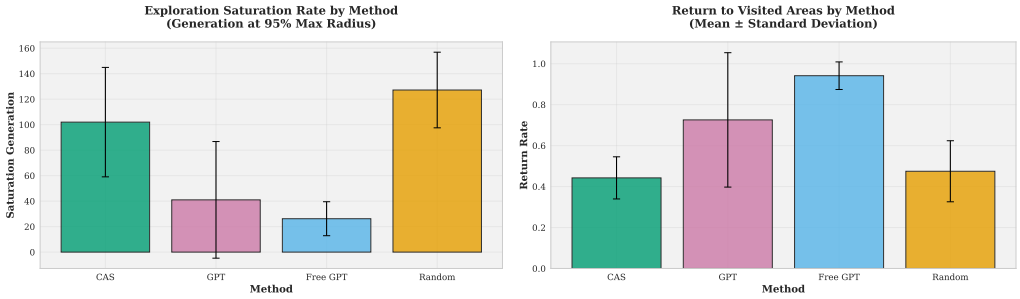
(a) (b)

Figure 22: (Left) Top 15 most frequently repeated concepts across runs for the GPT inspiration method. (Right) Top 15 most frequently repeated concepts across runs for Free GPT. The notation (+N) next to a concept indicates the number of additional distinct concepts that also met the repetition threshold in relation to that concept.

A.4.4 Long-Range Exploration details

Return Rate Computation: For each generation $t > 1$, we calculate the minimum Euclidean distance between the current embedding and all previous embeddings in the trajectory. A "return" occurs when this minimum distance falls below a threshold set to the median step distance for that specific inspiration method (making the metric adaptive to each method's natural exploration scale). The return rate is then computed as the fraction of all generations (excluding the first) that qualify as returns, providing a measure of how frequently an agent revisits semantically similar regions relative to its typical exploration granularity.

Saturation Generation Computation: We first compute the exploration radius for each generation as the Euclidean distance from the starting embedding to the current generation's embedding. The saturation generation is defined as the first generation where this radius reaches or exceeds 95% of the maximum radius achieved throughout the entire trajectory. This metric captures how quickly an agent reaches its peak exploration distance, distinguishing between methods that rapidly expand to their boundaries versus those that gradually explore outward over many generations. Fast saturation typically correlates with high return rates, as methods that quickly reach their exploration boundaries subsequently engage in local search behavior, repeatedly revisiting familiar semantic regions within their established trajectory.



(a) (b)

Figure 23: **Exploration behavior across agent types**: Saturation analysis (a) shows distinct strategies: CAS agents sustain exploration with late saturation (100 generations), Free GPT saturates fastest (25) then focuses on local search, Standard GPT is intermediate (40), and Random agents saturate slowest (130) due to random wandering. Return rate analysis (b) aligns with these patterns: Free GPT revisits most (0.95), CAS and Random are lower (0.45–0.48) due to more diverse conceptual exploration, while GPT is moderate (0.72). Together, the metrics show fast saturation coincides with high revisitation, reflecting local search after boundaries are reached.

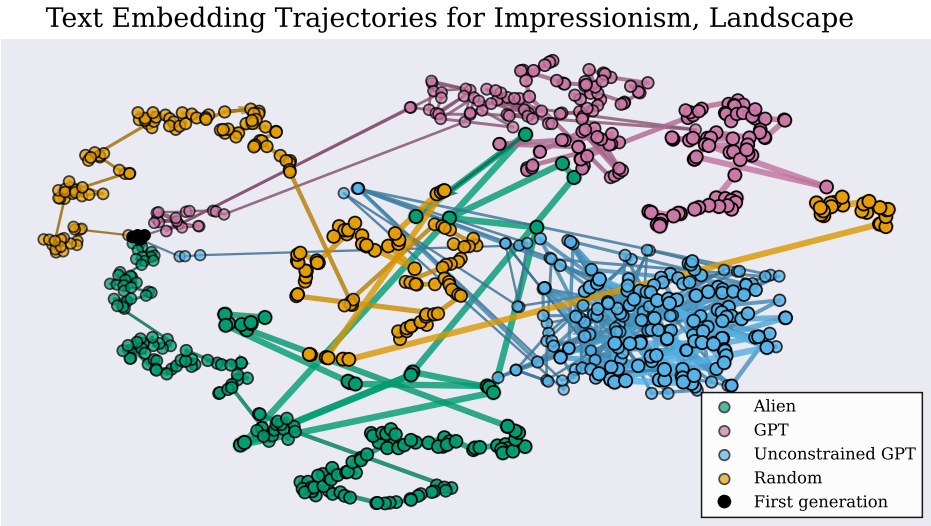


Figure 24: Example of 200 generations per method for starting concepts *Impressionism, Landscape*, showing trajectories in 2D t-SNE space by embedding each artwork's prompt. Larger points and wider edges indicate later generations.