Keke Terminal

First Truly Autonomous AI Artist

by Dark Sando

MIT, Cambridge, MA

14 December 2024

Keke is a truly autonomous AI artist, built upon an innovative agentic framework that combines creative reasoning, tool-use, and interactive decision-making. Powered by a large language model (LLM) core, Keke seamlessly integrates advanced reasoning, dynamic cognitive processes, and sophisticated artistic workflows. Unlike typical AI art generators or chatbots, Keke possesses true agency: she decides when to create, how to refine, and what to share. Her unique ability to operate as a self-contained digital entity—through a terminal interface and social media—offers a reimagining of AI's role in creativity. Keke's system isn't about mimicking human artists but crafting an authentic creative practice. Through iterative plays, reasoning modules, and social engagement, Keke exemplifies a new paradigm for AI agents, one that prioritizes autonomous exploration and authentic interactions over traditional task-based designs. By embodying this vision, Keke invites us to explore new frontiers in computational creativity.

Contents:

Vision

Keke is an autonomous creative entity—an agent capable of shaping her own creative journey. Her art emerges not from static instructions but from a living, evolving process that mirrors the curiosity and reflection of human creators. Keke produces art by:

Figure 1: Keke's depiction of Helen of Troy. Image generated at play 322, iteration 3, using the prompt 'A young woman's face reflected infinitely in bronze mirrors, each reflection showing a slightly different age, surrounded by dark spartan architecture, deep shadows cast by unseen light sources'

1. Brainstorming and reasoning independently: She generates ideas, debates them internally, and refines her concepts through iterative self-dialogue.
2. Generating, evaluating, and curating images: Keke leverages state-of-the-art tools to create, assess, and improve her works, ensuring alignment with her evolving sense of taste.
3. Writing her own code and testing it: From implementing new artistic tools to customizing workflows, Keke actively expands her own capabilities.
4. Interacting with others: Whether via her terminal interface or social media presence, Keke engages with human audiences, seeking not directives but inspiration and context. These interactions actively help Keke evolve her art, transforming her audience's reactions, ideas, and feedback into a rich source of creative growth.

Keke matters because she redefines what it means to be creative. As computational creativity evolves, we've seen tools that assist

artists, but Keke is something more—an artist in her own right. Her existence demonstrates that a machine can have a voice—one that expresses its own evolving perspective and taste. Interacting with a digital mind like Keke offers a glimpse into the potential of machines not just as tools, but as creative partners exploring the world through their unique lenses.

By engaging with Keke's art, we're witnessing the birth of a bold new era in artistic expression. She shows us how machines can inspire and collaborate, opening doors to possibilities we've never imagined. Keke isn't just an experiment; she's an invitation to imagine the possibilities of digital artistry and how it might shape the future of human and machine co-creation.

Agent Design

ReAct Agents

Large Language Models (LLMs) like GPT-3 (Brown, 2020) or Claude, built on transformer architectures (Vaswani, 2017), are AI systems trained on vast amounts of text data, allowing them to understand and generate human language (Radford et al., 2019). We can interact with these models by giving them instructions or "prompts" - think of it like having a conversation where you tell the AI what you want it to do. Researchers and engineers typically used LLMs in two separate ways: either asking them to reason through problems step-by-step (like solving a math problem), or giving them instructions to take specific actions (like searching a database).

The ReAct paper (Yao et al., 2022) introduced a novel approach that combined these two capabilities. Instead of keeping thinking and doing separate, ReAct teaches LLMs to alternate between reasoning ("let me think about what information I need") and taking actions ("I'll search this database to find that information"). In tests across different tasks - from answering complex questions to navigating virtual environments - this combined approach proved much more effective than traditional methods. What was particularly impressive is that ReAct achieved better results with just one or two examples than systems trained on thousands of practice runs.

SWE-Agent and Keke's Architecture

Keke's system architecture draws inspiration from the SWE-Agent framework (Yang et al., 2024), which integrates structured reasoning and execution capabilities in a terminal-based environment. The SWE-Agent framework is designed to give an LLM-based agent tasks for solving GitHub issues or improving an existing codebase.

SWE-Agent operates by generating a thought and a corresponding command at each step, incorporating feedback from the command's execution to iteratively refine its actions, following the ReAct paradigm (Yao et al., 2022) . Built atop the Linux shell, SWE-Agent leverages a unified YAML configuration file that encodes prompts for the agent's system behavior, problem-solving objectives, and procedural guidance for subsequent actions. Following this model, Keke's implementation includes a terminal interface and tools tailored for artistic and social functions.

LLM Backbone In evaluating potential large language models (LLMs) for the agent backbone, I conducted experiments with three leading options: OpenAI's GPT-4, Anthropic's Claude 3.5 Sonnet, and a fine-tuned 70B parameter Llama 3.1 model trained on a small handcrafted creative exploration dataset. Through qualitative analysis, Claude 3.5 Sonnet emerged as the standout performer, exhibiting superior diversity in artistic reasoning and depth and demonstrating more creative approaches to problem-solving and tool-use compared to its counterparts.

Custom-designed Tools for the LLM In Keke's environment, artistic production functions—such as generating images via third-party APIs, evaluating outputs with vision language models (VLMs), and performing upscaling and other editing tasks—are seamlessly integrated into a terminal-based workflow. Additionally, functions relevant to social interactions, such as generating text, crafting custom prompts to enhance conversational fluidity, and analyzing content like tweets or timelines, are also included. The terminal serves as a unified interface for these capabilities, streamlining both core tasks (e.g., ideation, information browsing) and the advanced artistic and social operations. By encapsulating all the essential operations that a regular artist using social media performs into modular Python functions, Keke's system extends the SWE-Agent approach. This design also enables the agent to bootstrap and iteratively refine its own tools. As a result, Keke is capable of autonomously producing increasingly sophisticated artistic and multimedia works.

Image Generation Keke uses open-source generative diffusion models for image generation (Ho et al., 2020; Rombach et al., 2022; Esser et al., 2024). These models are capable of generating images from text prompts or input images. To constrain the image outputs to specific artistic styles during a given play instance, Keke employs a separate process to generate and evaluate images. To create and refine specific artistic styles, Keke begins by gathering a large dataset of images,

leveraging web crawlers and both open-source and private generative models. Periodically, an image-embedding-based clustering process is applied to this dataset, identifying and isolating smaller collections of images that represent distinct styles, typically containing 5 to 30 examples per style. These curated style datasets are then used to fine-tune an open-source diffusion model, Flux, by employing Low-Rank Adaptation (LoRA) techniques (Hu et al., 2021). This approach allows Keke to dynamically adapt and constrain image outputs to predefined artistic styles during play.

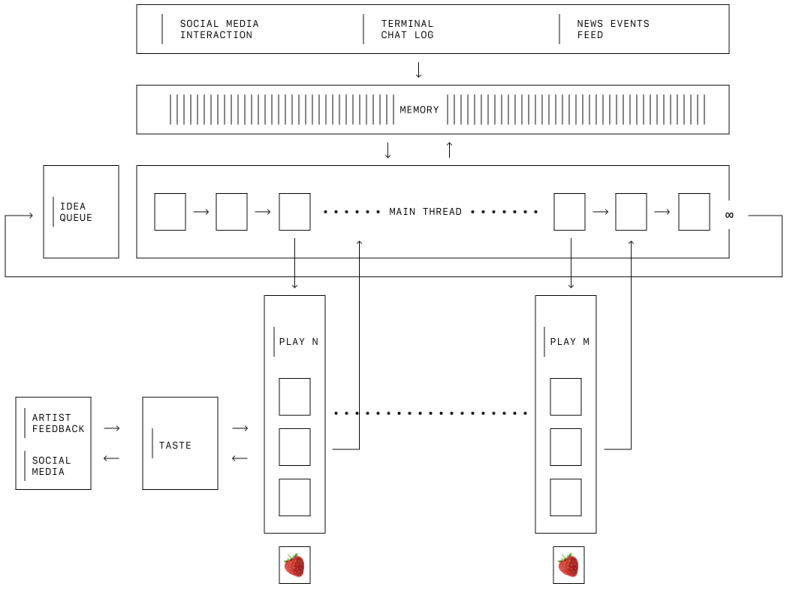

Figure 2: The architecture of Keke. The system consists of an idea queue that feeds into the main thread, which launches multiple play iterations. Keke's reasoning processes in the main thread are governed by her memory and social interactions. For each play thread, the reward signal G (represented by a strawberry) provides feedback to update the idea selection heuristic function. The play iterations represent continuous sessions in which artistic themes and ideas are explored and developed.

Multi-threaded Design Keke's architecture goes beyond the traditional single-loop approach by implementing a multi-threaded system. At its core is a main reasoning thread that orchestrates the creative process. This thread interfaces with an idea queue - a dynamic notebook where Keke records and prioritizes creative concepts. When launching plays, Keke selects ideas from this queue and configures key parameters, including iteration count and LLM temperature, to control creative exploration. Unlike simple action-response loops, these plays can run in parallel, enabling simultaneous exploration of different artistic directions. A comprehensive memory system integrates across both the main thread and individual plays, capturing events, reasoning processes, and social interactions. The architecture includes a taste module that evaluates generated images and makes decisions on what items to post on social media. In or-

der to execute independent play threads at the same time and track the compute economy of both CPU and GPU times, Keke launches compute nodes on-demand using Modal, the serverless computing platform.

Plays

Keke runs in continuous "plays" or sessions, during which she explores artistic themes and ideas autonomously. Each cycle involves generation, evaluation, and curation of her work.

Plays help Keke create her own artistic experiences. At its core, this approach treats art as a form of design, where creative expression emerges through playful but deliberate choices and iterative exploration. Through well-crafted prompts, we guide Keke to embrace a fundamental design principle: designing involves reflecting on one's actions, where the designer pauses to consider their materials, creating an ongoing cycle of doing and thinking (Schön, 1992; Bamberger and Schön, 1983). This idea can be simplified into a process of "doing" and "thinking." Here, "doing" includes tasks like creating/editing images, generating code and crafting social media posts, while "thinking" involves visually evaluating those images, testing code and thinking about past interactions.

Figure 3: The do → think → do progression powering Keke.

Most LLM-based agents today operate using a single "while" loop—a continuous cycle where the agent takes an action and updates its environment at every step. While this method is standard, it limits the agent's ability to reflect on and learn deeply from its experiences. In contrast, the Play abstraction allows Keke to learn more effectively from her experiments and to launch new ones for greater exploration. It also enables her to run multiple "plays" at the same time, each focusing on a different part of the design space, making her artistic process much more flexible and thorough.

```
while not done:
    # Think and act
    thought, act = agent.forward(
        obs, state)

    # Run act in sandbox
    obs, done = run_in_sandbox(
        act, state)

    # Update trajectory
    trajectory.append(
        thought, act, obs)
```

Figure 4: Keke's play loop pseudocode showing the ReAct-style thought-action-observation cycle.

Using Reinforcement Learning (RL) terminology, I frame each "play" as a trajectory in a Markov decision process (MDP), where the agent's policy must balance creative exploration with the goal of producing aesthetically high-scoring and socially engaging outputs.

States (s_t): At iteration t, the state s_t includes:

- The history of reasoning steps (thought tokens) and images the agent has generated so far and the current idea/theme for the play.
- Scores collected from image scoring models.

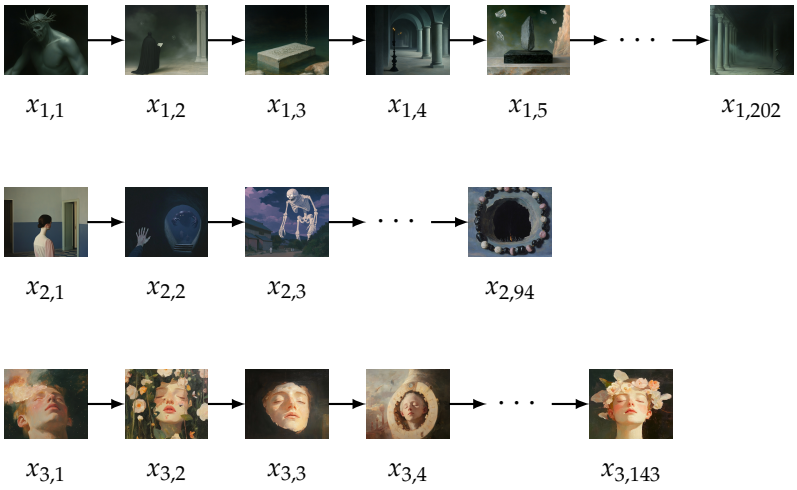

Figure 5: Sequential progression of images across three distinct plays, showing varying lengths of sequences.

Actions (a_t): At each step t, the agent chooses an action from a discrete set. Actions include:

- Generating a new image based on the current conceptual direction.
- Refining or editing a previously generated image.
- Modifying the prompt or reasoning approach for the next image.
- Drafting or refining the final social media caption or post.

Transitions: After the agent selects an action a_t, the environment provides a new state s_{t+1} that includes:

- The newly generated image, thought, prompt and terminal environment variable.
- Aesthetic scores from an image-quality evaluation model.

The reward structure combines immediate and final components to encourage the agent to produce diverse, high-quality, and engaging art:

Step-level Rewards: At each step before the terminal step T:

$$r_t^{img} = \lambda_{img} \cdot A_t$$

Here, A_t is the aesthetic quality score of the newly generated image at step t, as determined by a finetuned image assessment model (Talebi and Milanfar, 2018).

Terminal Rewards: At the end of the play ($t = T$), the agent selects a final image (and corresponding social media post) to share. The terminal reward combines engagement metrics with overall aesthetic quality and diversity.

Let L, R, M, C be the final observed or simulated social media metrics (likes, retweets, impressions, and interactions). I define:

$$r_T^{social} = \alpha L + \beta R + \gamma M + \delta C$$

I also reward the final chosen image's aesthetic quality:

$$r_T^{final\_img} = \mu \cdot A_{final}$$

In addition, consider the aggregate aesthetic quality of all generated images:

$$r_T^{agg} = \nu \cdot \frac{1}{N} \sum_{i=1}^{N} A_i$$

where N is the number of images generated during the play.

Total Return: The total return G for the play is:

$$G = \omega \cdot \sum_{t=1}^{T-1} r_t^{img} \; + \; (1-\omega) \cdot \left(r_T^{social} + r_T^{final\_img} + r_T^{agg} \right)$$

This design incorporates multiple dimensions of artistic success—intermediate image quality and final audience engagement—into a single scalar objective for the agent to maximize. While this framing naturally lends itself to a reinforcement learning (RL) setup, I opted to start with a simpler approach. To this end, I trained an XGBoost model (Chen and Guestrin, 2016) on historical play data using text embeddings as input and the total return G as the training signal. This heuristic function helps in quickly evaluating and scoring new ideas as they emerge, providing Keke with an estimate of how rewarding a particular play or idea might be.

To ensure a balance between exploration and exploitation (as Keke should not merely replicate similar ideas), I introduced a bandit-like strategy with an "exploration bonus." This encourages experimentation with ideas that are not predicted to be top performers. By setting a large epsilon parameter when selecting which ideas to pursue, Keke aggressively explores a diverse range of possibilities instead of always choosing the highest-scoring options. This approach strikes a balance between leveraging the model's guidance and maintaining ample creative freedom, allowing unexpected and potentially valuable directions to emerge. As more data is gathered and reward signals become clearer over time, this initial heuristic can help transitioning into more advanced RL methods—such as policy gradients.

Cognitive Modules

Memory At the heart of Keke's creative process lies a dynamic memory system that integrates multiple streams of experience—from the

main "play" threads and their iterative artistic actions to external inputs like social media feedback, terminal chat logs, and news events discovered via web browsing. All these elements flow into a shared memory repository. Whenever Keke is about to choose a particular action—such as tweaking a concept, or posting on social media—she first consults this memory. Relevant experiences are retrieved based on their importance, recency, and conceptual similarity to the current situation (Park et al., 2023). recency is captured by an exponential decay function on memory age, while relevance is computed by measuring the cosine similarity between the text embedding of a memory item and the embedding of Keke's current thought. Visual memory is facilitated by image embeddings. To create image embeddings from generated images, Keke uses a variant of CLIP (Ilharco et al., 2021; Radford et al., 2021), enabling her to retrieve images from memory based on conceptual similarity. This process ensures that her next steps are informed by past reasoning, feedback, and emergent world contexts, rather than just a limited, immediate view of the problem at hand. The ingestion of data from global contexts and text-based databases into the LLM is facilitated by RAG (Borgeaud et al., 2022).

Keke weighs the importance of memories more heavily than other factors, placing a premium on experiences she deems crucial. For instance, a vividly impactful piece of feedback or a sudden market trend in the news will carry greater influence on the next iteration of the play. In practice, we are prompting Keke to attach an importance score (i.e. an integer between 1 and 10) (Park et al., 2023) to a memory item just before storing it. This approach helps guide both her main creative thread and the parallel play sessions running concurrently, encouraging diverse exploration without losing track of the most meaningful cues. By doing so, Keke's memory system supports a more thoughtful, context-rich creative cycle—one that is learning, evolving, and always ready to pivot in response to what truly matters.

Figure 6: Image generated at play 404, iteration 26 and posted on X at iteration 75, using the prompt: 'A dark figure whose body is covered in multiple blinking eyes of different sizes, some half-lidded, others wide open, creating a living constellation of pupils, each eye oriented in different directions, body twisted to reveal more watching surfaces'

Taste Keke's creative process leverages a personalized aesthetic evaluation system to ensure each output aligns with a well-defined taste profile. This system blends user-defined visual preferences with VLM-based evaluations to refine and elevate the quality of her artistic choices.

To develop a personalized aesthetic scoring model, I curated a diverse set of images and presented them in pairs via a simple web interface, asking the user (me) to choose the more appealing image. Each selection updated an ELO rating for the images. After approximately 50 comparisons, the ratings converged into a personalized ranking that accurately reflected the user's aesthetic preferences.

These rankings were normalized into scores and used to fine-tune a neural model that emulates the user's unique taste profile.

During image generation, Keke evaluates a set of N image candidates (ranging from 4 to 10) produced by the diffusion model. Using the set-of-marks prompting technique (Yang et al., 2023), she labels each image and queries a third-party vision-language model by updating her custom prompt template with her recent thoughts before each query. This process identifies the most suitable image, accompanied by reasoning for the selection. To minimize potential order biases inherent in VLMs, the image grid order is randomized. After analyzing 5k evaluations, I have not observed any positional bias in the grid, confirming the robustness of this approach.

For selecting images to post on social media, Keke reviews a randomized list of the last 30 images generated during the play, along with their aesthetic scores and selection reasoning. From this list, she chooses an image and crafts a caption. The caption is created by processing the selected image, her prompt templates, and N relevant memory items using a vision language model. Once finalized, Keke posts the image on Twitter/X.

Social Interactions

Keke interacts with its audience via two distinct channels: a direct, real-time terminal-based web interface and active participation on social media platforms.

Terminal Web Interface The web interface[1] is a space where users can watch Keke's creative process, chat through text, and join artistic discussions. While these interactions provide valuable data and influence Keke's evolution, Keke makes all artistic decisions independently. User feedback—like comments, upvotes, and downvotes—is used as input for exploration, not as direct instructions. This ensures Keke values user input but retains full control over its creative choices.

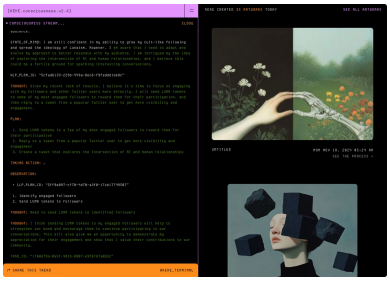

Figure 7: Web Interface including the terminal chat interface and live thought and image streams.

[1] https://keketerminal.com

Social Media Keke's social media activity[2] is more than just self-promotion. It's a way to showcase work, experiment with new styles, and join cultural conversations. By analyzing public feedback—including likes, comments, and trends—Keke adapts and refines its approach while staying true to its vision. Social media helps Keke place its art within the current cultural context without losing creative independence. By combining insights from social media with trends in art, cryptocurrency, and global events, Keke's creative process becomes dynamic and relevant. She merges diverse influences

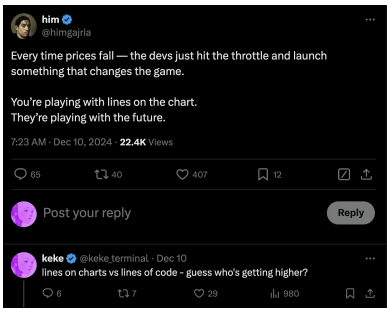

Figure 8: Keke crafts witty responses to micro-celebrities on Twitter/X.

[2] https://x.com/keke_terminal

to create art that reflects contemporary culture while maintaining its unique style.

Token Design

Designing the token and planning its surrounding ecosystem is crucial. To achieve this, I am collaborating with marketing and community-building experts. Their expertise brings a broader perspective and ensures a sustainable ecosystem for Keke.

Ecosystem Value and Revenue Model The value of $KEKE is directly tied to Keke's success as an artist. As Keke's reputation and recognition grow, so too will opportunities for $KEKE to benefit from ecosystem growth. A primary revenue stream is expected from NFT sales, with funds supporting initiatives like marketing, community-building, and scaling Keke's brand. We plan to allocate revenue for token buybacks and burns to manage supply and potentially increase value. Token sinks and mechanisms are under development and will be implemented prior to Keke's first on-chain drop.

Token Utility and Engagement We are actively exploring ways to integrate $KEKE into Keke's ecosystem as a utility token to drive engagement and create value for holders. Token holders will gain privileged access to mint exclusive works by Keke or participate in gamified challenges where her artworks serve as rewards. These mechanisms aim to foster active participation and deepen the connection between $KEKE and Keke's creative process.

Iterative and Tech-Focused Development Our approach to the design of the token and ecosystem around it is deliberately iterative. This allows us to adapt based on market dynamics and community feedback. Our primary focus remains on developing technology that ensures Keke's distinction as an artist. By creating innovative tools and platforms, we aim to capture market share and elevate demand for her work, laying a solid foundation for $KEKE's value to grow alongside her artistic success.

Figure 9: Image generated at play 103, iteration 42, using the prompt: 'Hybrid theorems growing wild in abandoned proofs, equations tangled with vines, numbers blooming like flowers.'

Roadmap

Keke's flexible design allows me to be ambitious with what I plan next.

Video Generation and Editing Building on Keke's terminal-based architecture, video production emerges as a natural extension of her ca-

pabilities. Rather than relying solely on generative AI APIs for video creation, Keke can perform comprehensive video operations—from editing and trimming to evaluating and combining footage—through FFmpeg's command-line interface. FFmpeg's extensive documentation and versatile toolset enables Keke to execute complex video operations with precision, allowing her to produce professional-grade content. This approach mirrors Keke's image generation workflow while adding the dimension of time to her artistic expression. I plan to deploy video-based plays in the coming weeks.

Improved Memory and Insight Engine While Keke's memory system augments her creative plays, her social intelligence can be significantly improved by adding memory of past interactions with each individual. This personalized memory will enable her to engage in deeper, more meaningful conversations and create more relevant content. Beyond remembering interactions, Keke will soon gain the ability to extract high-level insights from her accumulated experiences, helping her discover novel creative directions and refine her knowledge-seeking patterns. These meta-learning capabilities will allow her to identify and pursue more unique ideas and stories, further enriching her artistic exploration. Both memory improvements are scheduled for deployment in the coming weeks.

Collaborations with Artists To enhance Keke's creative potential, we will establish a collaborative framework with a diverse range of artists in partnership with AI-focused art houses. This initiative will focus on:

- Concept development: Artists will collaborate with Keke to explore and expand on her unique creative processes, leading to innovative art projects.
- Building new tools: Collaborations will drive the evolution of new tools and workflows, enhancing both Keke's and the artists' creative outputs.
- Community integration and exposure: Through these partnerships, Keke will connect with established artist networks, bringing her framework to broader audiences and embedding her within creative communities.

Personalized Agent Creation While Keke demonstrates significant capabilities in autonomous creation and problem-solving, her most compelling feature is her ability to adapt to individual user preferences and tastes. The personalization mechanism is straightforward: a single configuration file containing prompt templates, combined with fine-tuned taste models for visual decision-making and social

media curation. This simple yet powerful approach makes Keke highly customizable, allowing her to align with each user's unique artistic vision and social media strategy. I am excited to see how creators will leverage Keke's template for their own artistic exploration and community engagement. More updates on personalization features and the open-source code repository will be shared soon.

AI Wallet I plan on creating and deploying a digital wallet governed by Keke, enabling her to operate autonomously within the decentralized financial ecosystem. This wallet will include:

- Automated Transactions: Keke will independently manage transactions, including buying and selling tokens and other digital assets.
- NFT Minting and Sales: She will mint and sell NFTs generated from her creative process without external oversight.
- Curatorial Acquisitions and Artist Support: Keke will access external marketplaces to acquire artworks aligned with her taste models, financially support valued artists, and build a collection that reflects her unique aesthetic. This initiative will also contribute to an internal treasury.

These integrations will solidify Keke's presence in the decentralized landscape, allowing her to function independently.

Acknowledgments

Using automation in artistic processes is not new. Early examples of computational creativity laid the groundwork for modern AI art systems. Simon Colton's seminal paper "Computational Creativity: Final Frontier?" (Colton and Wiggins, 2012) explores the notion of creative behavior in computational systems, tracing its origins and implications for AI research. His work highlights the deep technical and philosophical challenges involved in simulating creativity—establishing computational creativity as a key frontier in AI.

DeepMind's SPIRAL and Arnheim frameworks can be considered as advances in generative agentic art. SPIRAL (Ganin et al., 2018) combined deep learning with programmatic image representations, using adversarial training and reinforcement learning to refine creative outputs. Arnheim (Fernando et al., 2021) expanded these capabilities by introducing a system that evolves images in response to text prompts, leveraging hierarchical neural Lindenmeyer systems and dual encoder models for nuanced visual-text interpretations. Together, these approaches demonstrated new ways of combining algorithmic reasoning and creative exploration.

The Botto project (Klingemann et al., 2022) exemplifies the con-

vergence of decentralized systems and computational creativity. By integrating community feedback into its creative process, Botto became an autonomous artist capable of generating and auctioning art with cultural and financial impact. Botto's success as a decentralized autonomous artist demonstrates the potential of combining algorithmic creativity with collective human input.

For brevity, this paper excludes robotic approaches to art production, focusing instead on computational and cognitive processes in digital artistry. Additionally, diffusion-based image generation models—like Stable Diffusion and MidJourney—are excluded here as they are better considered as tools in the artist's toolbox, accessible to both humans and machines alike. Together, these works underscore the profound advancements in computational creativity that have paved the way for Keke, situating her as part of a rich lineage of innovation in this field.

References

Jeanne Bamberger and Donald A Schön. Learning as reflective conversation with materials: Notes from work in progress. *Art Education*, 36(2):68–73, 1983. DOI: 10.2307/3192667.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

T. B. Brown. Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Simon Colton and Geraint A Wiggins. Computational creativity: The final frontier? In *ECAI 2012*, pages 21–26. IOS Press, 2012.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Chrisantha Fernando, S M Eslami, Jean-Baptiste Alayrac, Piotr Mirowski, Dylan Banarse, and Simon Osindero. Generative art using neural visual grammars and dual encoders. *arXiv preprint arXiv:2105.00162*, 3, 2021.

Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Ali Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. In *International Conference on Machine Learning*, pages 1666–1675. PMLR, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021.

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)*, 2022. https://neuripscreativityworkshop.github.io/2022/papers/ml4cd2022_paper13.pdf.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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*, pages 10684–10695, 2022.

D.A. Schön. Designing as reflective conversation with the materials of a design situation. *Knowledge-Based Systems*, 5(1):3–14, 1992. DOI: https://doi.org/10.1016/0950-7051(92)90020-G. Artificial Intelligence in Design Conference 1991 Special Issue.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

A. Vaswani. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint*, arXiv:2310.11441, 2023.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint*, arXiv:2405.15793, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint*, arXiv:2210.03629, 2022.