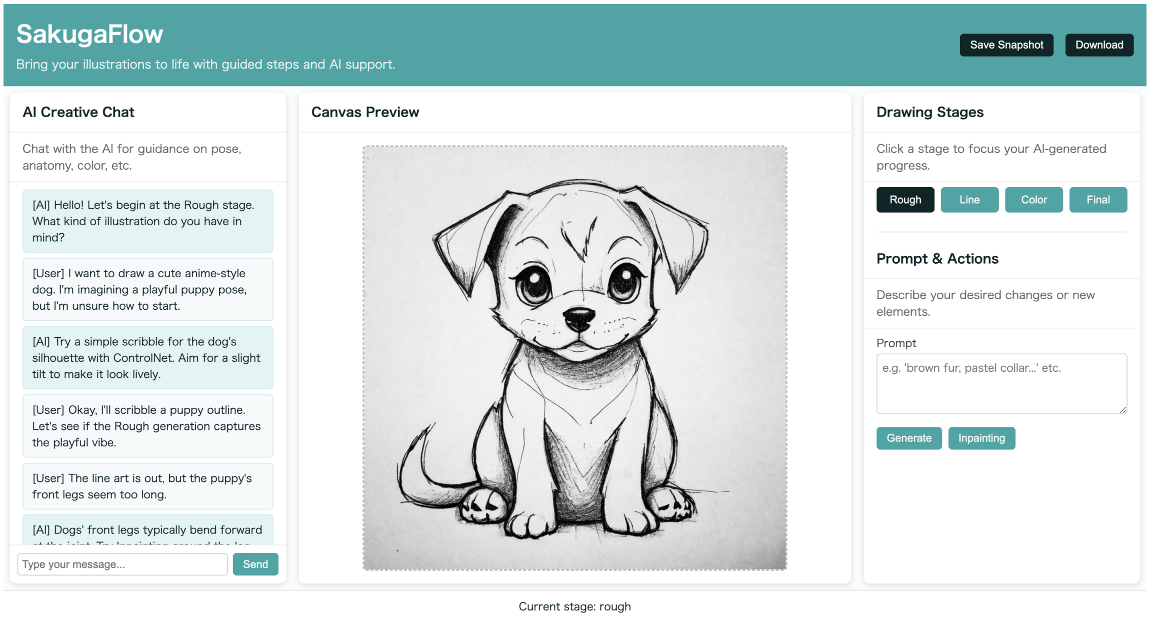
SakugaFlow: A Stagewise Illustration Framework Emulating the Human Drawing Process and Providing Interactive Tutoring for Novice Drawing Skills

ANONYMOUS AUTHOR(S)



SakugaFlow

Bring your illustrations to life with guided steps and AI support.

Chat with the AI for guidance on pose, anatomy, color, etc.

[AI] Hello! Let's begin at the Rough stage. What kind of illustration do you have in mind?

[User] I want to draw a cute anime-style dog. I'm imagining a playful puppy pose, but I'm unsure how to start.

[AI] Try a simple scribble for the dog's silhouette with ControlNet. Aim for a slight tilt to make it look lively.

[User] Okay, I'll scribble a puppy outline. Let's see if the Rough generation captures the playful vibe.

[User] The line art is out, but the puppy's front legs seem too long.

[AI] Dogs' front legs typically bend forward

Click a stage to focus your AI-generated progress.

Describe your desired changes or new elements.

Prompt: e.g. 'brown fur, pastel collar...' etc.

Current stage: rough

Fig. 1. Overview of SakugaFlow's UI and architecture. (1) Users progress from rough to line to color to finish, each powered by a diffusion-based backend. (2) An LLM tutor provides real-time feedback on anatomy, perspective, and more. (3) Users can branch versions, inpaint local regions, and compare alternatives within the same project.

While current AI illustration tools can generate high-quality images from text prompts, they do not explain how a person could replicate the final drawing. We present **SakugaFlow**, a tool that gradually refines an illustration through four key stages—rough sketch, line art, coloring, and final finishing—and allows users to revise at any point along the way. SakugaFlow integrates diffusion-based image generation at each stage with an AI agent that can advise and consult on how to proceed at the next step. Powered by a Large Language Model, this agent helps novice users understand foundational principles such as anatomy and composition, and offers localized revision guidance. By revealing intermediate outputs at each stage, our system transforms generative AI from a black-box producer into a *learning partner*, enabling beginners to develop their skills as they iteratively refine their artwork.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Collaborative content creation*; • **Applied computing** → *Interactive learning environments*; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: Generative AI; Educational Dialogue Systems; Intelligent Tutoring Systems; Human-AI Co-Creation

Manuscript submitted to ACM

ACM Reference Format:

Anonymous Author(s). 2018. SakugaFlow: A Stagewise Illustration Framework Emulating the Human Drawing Process and Providing Interactive Tutoring for Novice Drawing Skills. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Generative models for illustration have advanced rapidly, with diffusion-based approaches now producing remarkably detailed images from simple prompts [8, 11]. However, these one-shot techniques rarely show the *drawing process* that underlies the final output. For a beginner hoping to improve their own skills, seeing only a completed image provides little guidance on *how* lines, colors, or forms evolve.

Human artists, by contrast, usually begin with a loose rough sketch, transition to cleaner line art, apply broad color blocks, and then refine shading and details. Each step can be a learning opportunity: understanding silhouette, proportion, color harmony, and so forth. If an AI system bypasses—or at least does not reveal—these steps, novices miss the gradual refinement process that connects simple fundamentals to complex finished pieces.

SakugaFlow addresses this gap. Rather than presenting an immediate final result, our system divides the illustration pipeline into four explicit stages: rough sketch, line art, coloring, and final finishing. At each stage, users see a partially complete image generated by a diffusion model, which they can revise or question with the help of an educational dialogue agent. This agent, built on a Large Language Model (LLM), offers targeted explanations about anatomy, perspective, and design principles. By directly associating each incremental change with an interpretable step, SakugaFlow transforms a black-box generator into a *scaffolded learning environment*, where users not only observe but also *actively influence* how the image develops.

Contributions. Our key contributions are as follows:

- **Staged diffusion aligned with human drawing steps.** While prior diffusion methods mainly offer localized control of final outputs, we re-map them into four phases: rough sketch, line art, coloring, and finishing. This approach clarifies *how* the final image emerges and allows novices to experiment with each step.
- **Integration of an educational dialogue agent for real-time feedback.** Although interactive tutoring systems have seen success in text-based domains, their application to drawing is underexplored. Our system incorporates an LLM to connect each generative step with relevant art principles, offering immediate explanations (e.g., perspective corrections) and adaptive guidance throughout the process.
- **Fusing creative assistance with skill development.** Unlike prior tools focused on final output or user creativity, we emphasize helping beginners build foundational illustration skills. By positioning AI as both teacher and partner, we enable stepwise practice, explorative feedback, and reflective learning—prioritizing skill acquisition over mere production efficiency.

2 Related Work

AI-based Illustration Generation: Focus on Diffusion Models. Diffusion models [8, 11] have gained prominence as a powerful alternative to GANs [5, 9, 10, 16], enabling high-quality image generation through iterative noise removal. Extensions like ControlNet, Prompt-to-Prompt, and Inpainting [7, 15] allow refined control over line art, pose specification, and text-based editing, while Promptify [1] similarly offers interactive prompt exploration using LLM. However, these existing methods focus primarily on the creation or editing of the finished image, and do not model the drawing process of how humans complete their drawing from nothing, nor are they able to extract such information.

Illustration Learning Support and Process Visualization. Tutorial videos and step-by-step guides are widely available, but they often end up being passive viewing, and it is difficult to choose the learning material you want. Logging or replay of drawing processes [4, 6] can show expert workflows via timelapse but typically lack real-time user input for revisions or questions. Real-time drawing aids [14] focus on hand-drawn input, such as automatic error correction or feedback, but have not yet explored the perspective of *learning from AI-generated* intermediate outputs.
Creative Support and Human-AI Co-Creation. Recent commercial software, such as Photoshop (Generative Fill) and Illustrator (Generative Recolor), integrates AI for quick image and vector transformations [12]. Academic studies likewise investigate real-time AI collaboration in drawing [2, 3] and prompt-based image generation [1, 13]. While many of these systems boost final-image efficiency or assist with ideation, they rarely address how novices *learn* illustration step by step. We propose a stage-based AI approach combined with educational dialogues, allowing beginners to observe each process phase—rough sketch, line art, coloring, and finalizing—and receive guided support. By focusing on progressive skill acquisition rather than just output optimization, our work fills an important gap in existing co-creation research.

3 Proposed System: SakugaFlow

3.1 Design Goals and Rationale

SakugaFlow aims to help beginners learn how illustrations progress through key stages rather than jumping directly to a final image. Human artists typically do a rough sketch, refine line art, block colors, then add finishing details. We integrate a diffusion model into these steps so novices can see partial outputs and adjust them step by step. This structure ties closely to educational needs, providing a guided, interpretable journey through each stage of the creative process.

3.2 System Overview and UI

As shown in Fig.1, SakugaFlow presents a four-phase workflow:

(1) **Rough Sketch:** Users define basic shapes or a scribble-based layout (optionally with ControlNet).
(2) **Line Art:** A Prompt-to-Prompt transformation clarifies contours; Inpainting can refine errors (e.g., proportions).
(3) **Coloring:** The system suggests color palettes or shading styles; users can branch multiple color studies.
(4) **Finishing:** Final highlights, lighting, and minor adjustments lead to the completed illustration.

A chat pane (right or left side) connects to an LLM-based "tutor," which offers quick theory tips (e.g., "Why add line thickness here?"). Users can also request partial re-generation by selecting areas to inpaint or rewriting the prompt. A branching manager lets them save and compare multiple versions of the same stage.

3.3 Interaction Flow

1) Project Setup. The user inputs a theme (e.g., "fantasy character"), then starts at the Rough stage. **2) Rough Stage.** ControlNet(*Scribble*) or a blank canvas yields a rough shape. The LLM tutor might suggest adjusting pose or composition. **3) Line Art Stage.** Prompt-to-Prompt refines the lines; users correct details (e.g., mismatched arm length) via Inpainting. **4) Color Stage.** The system proposes palettes; a user can branch to compare two coloring approaches. The tutor explains color theory (e.g., warm vs. cool contrast). **5) Finishing.** Final touches (lighting, shading). The tutor checks consistency and prompts reflection ("Where is the light source?").

Users can backtrack or create alternative branches. *Active experimentation* is encouraged rather than showing a single final image.

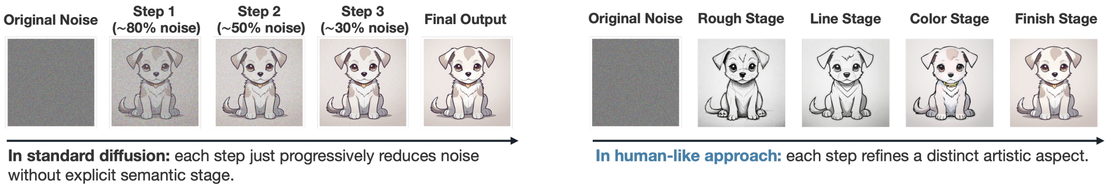

Original Noise
Step 1 (~80% noise)
Step 2 (~50% noise)
Step 3 (~30% noise)
Final Output
In standard diffusion: each step just progressively reduces noise without explicit semantic stage.
Original Noise
Rough Stage
Line Stage
Color Stage
Finish Stage
In human-like approach: each step refines a distinct artistic aspect.

Fig. 2. **A human-like stepwise approach (left) vs. a standard diffusion process (right).** Top: the illustration evolves through rough, line, color, and final stages. Bottom: the model gradually removes noise (e.g., $\sim 80\% \longrightarrow 0\%$) without clear drawing phases.

3.4 Implementation Details

Frontend. Built in React with a canvas for partial previews, an Inpainting brush, and a chat interface. Minimizing latency is crucial; we use a small in-browser aggregator for partial updates. **Backend.** Python (FastAPI) orchestrates:

- *Diffusion Model:* Stable Diffusion + ControlNet for each stage.
- *LLM (GPT):* Processes user queries and context from the current stage to generate relevant tips.
- *Versioning/Branches:* Stores user steps, prompts, and image states, enabling revert and compare.

All major computations run on a GPU server, returning images to the frontend asynchronously. Future optimizations include caching partial results for repeated local edits.

3.5 Optional Hand-Drawn Input

While novices can rely solely on text prompts and partial scribbles, advanced users may integrate their own detailed sketches to preserve personal style or practice motor skills. SakugaFlow can treat these sketches like “control images,” merging them with generated content.

4 Discussion & Future Work

Limitations and Future Directions. Although SakugaFlow introduces separate steps (rough, line, color, finish), our backend diffusion model remains optimized for final outputs, limiting the interpretability of intermediate states. Future efforts include constructing more granular representations (e.g., partial sketches or layered data), training explicitly on sequential refinements, and running controlled user studies to quantify skill acquisition. Balancing fast, one-shot generation with slower, staged processes is also crucial for scenarios where iterative learning is desired.

Toward a Human-Like Staged Diffusion Approach While standard diffusion treats image generation as a continuous noise-removal sequence, a human-like staged approach aligns more closely with how artists actually work. Fig. **??** contrasts these methods: the top row shows discrete phases (rough, line, color, final), while the bottom row removes noise without revealing clear drawing steps. By making each phase explicit, we can offer finer-grained visualization of *when* lines, color, and details emerge, providing a more intuitive interaction for novices who naturally think, “draw lines first, then add color, then refine.” Meanwhile, these intermediate stages can be reused as modular sub-tasks (e.g., silhouette extraction, line analysis, color application), opening the door to new multi-task pipelines or partial retraining. Finally, segmenting the process in a human-like manner enhances interpretability and debugging: if a line stage is flawed, we know exactly where to intervene. This also helps in settings where safety or quality assurance requires each step to be validated before proceeding.

In conclusion, SakugaFlow represents an initial step toward interpretable, pedagogically rich generative illustration. Further aligning the underlying models with human workflows—and rigorously evaluating those alignment strategies—remains our primary goal for future work.

References

[1] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Article 96, 14 pages.

[2] Nicholas Davis, Chih-Pin Hsiao, Kunwar Yashraj Singh, and Brian Magerko. 2016. Co-creative drawing agent with object recognition. In *Proceedings of the Twelfth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'16)*. Article 2, 7 pages.

[3] Nicholas Davis and Janet Rafner. 2025. AI Drawing Partner: Co-Creative Drawing Agent and Research Platform to Model Co-Creation. arXiv:2501.06607 [cs.HC] https://arxiv.org/abs/2501.06607

[4] Jennifer Fernquist, Tovi Grossman, and George Fitzmaurice. 2011. Sketch-sketch revolution: an engaging tutorial system for guided sketching and application learning *(UIST '11)*. 373–382.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 27. 2672–2680.

[6] Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2010. Chronicle: capture, exploration, and playback of document workflow histories *(UIST '10)*. 143–152. doi:10.1145/1866029.1866054

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626* (2022).

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 6840–6851.

[9] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[10] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434* (2015).

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[12] Ian P. Swift and Debaleena Chattopadhyay. 2024. A Value-Oriented Investigation of Photoshop's Generative Fill. arXiv:2404.17781 [cs.HC] https://arxiv.org/abs/2404.17781

[13] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. 1–21. doi:10.1145/3613904.3642803

[14] Blake Williford. 2017. SketchTivity: Improving Creativity by Learning Sketching with an Intelligent Tutoring System. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition* (Singapore, Singapore) *(CC '17)*. 477–483.

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision (ICCV)* (2017).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009