FIGUR: An AI-Driven Collaborative Robotic System for Expressive Portrait Generation

Hojeong Kim
Department of Computer Science and Engineering
Sogang University
Seoul, South Korea
amy22@sogang.ac.kr

Jinho Son*
Department of Computer Science and Engineering
Sogang University
Seoul, South Korea
jinhoson@sogang.ac.kr

***Abstract*—This paper presents FIGUR (Face Image Generation Using Robot), an interactive AI-enhanced robotic portrait system that integrates Multimodal Large Language Models (MLLMs) with collaborative robotics to generate personalized drawings with expressive movements. Unlike conventional systems that produce purely functional movements, FIGUR employs a pipeline combining Mediapipe for face detection, morphological skeletonization algorithm and greedy nearest-neighbor heuristic, resulting in positive user satisfaction feedback. The system employs an Hanwha HCR-5 collaborative robot and implements expressive movement primitives inspired by the ELEGNT framework [1], including artistic flourish gestures. Real-time face detection via MediaPipe Face Mesh with frontal-gaze scoring ensures optimal input quality by selecting the top-ranked frames from 300 captured samples. Experimental results show an average drawing completion time of 2.4 minutes with high line accuracy. User feedback (N=32) indicates that theatrical flourish movements enhance user engagement and perceived robot creativity compared to functional movements alone. The system also demonstrates the ability to render distinct artistic styles while minimizing pen-travel distance through trajectory optimization.**

***Keywords*—Human-robot interaction, physical AI, expressive robotics, collaborative robots, multimodal LLM**

I. Introduction

The integration of artificial intelligence with collaborative robotics has opened new possibilities for creative human-robot interaction (HRI). While industrial robots excel at repetitive precision tasks, their application in artistic and social contexts remains limited by purely functional movement paradigms that lack expressive qualities. Traditional robotic drawing systems face two critical limitations: (1) edge-detection algorithms produce noisy, fragmented lineart requiring extensive post-processing, and (2) purely functional robot movements create mechanical, uninspiring user experiences that fail to convey artistic intention or engage users emotionally. This paper addresses these challenges through FIGUR, a comprehensive system that makes three primary contributions:

A. Theatrical Flourish Implementation

Drawing on principles from performing arts and the ELEGNT framework from Apple for non-anthropomorphic robot expressiveness, we implement post-drawing theatrical flourish gestures—sweeping circular movements that mimic a conductor's final hand stroke or a stage performer's presentation bow. User studies demonstrate that these artistic flourishes increase perceived robot creativity and user engagement compared to purely functional movements.

B. Semantic Abstraction

The integration of Gemini-2.5-Flash to transform webcam captures into single-stroke compatible line art, adhering to specific stylistic constraints.

C. Topological Vectorization

A shift from boundary tracing to morphological skeletonization, ensuring single-pixel width trajectories suitable for pen plotters.

Fig. 1. FIGUR embodiment and workspace

II. Related Work

A. Robotic Drawing and Creative Systems

Recent advances in robotic drawing systems have demonstrated the potential for AI-enhanced creative robotics. FRIDA (Framework and Robotics Initiative for Developing Arts),

developed at Carnegie Mellon University, pioneered the integration of large language models with robotic manipulation for artistic creation [2]. FRIDA uses a collaborative robot arm equipped with various artistic tools and employs vision-language models to translate textual prompts into painting actions. While FRIDA demonstrates impressive creative capabilities through iterative refinement and multi-modal feedback, it focuses primarily on abstract painting rather than portrait drawing and does not address expressive movement design. Scratch to Sketch introduced decoupled hierarchical reinforcement learning approach for robotic sketching [3]. However, like FRIDA, it treats robot movement purely functionally, optimizing for drawing accuracy without considering the expressive or social dimensions of human-robot interaction. In addition to painting-oriented systems, prior work has explored autonomous robotic portrait generation. Song et al. proposed a multi-stage portrait drawing system that integrates facial feature extraction with stroke-based rendering [4]. While highly optimized for visual accuracy, their robot motions remain purely functional without expressive elements. Compared to these previous works, FIGUR makes two distinct contributions: (1) integration of state-of-the-art MLLM for line art generation, producing cleaner, more artistic outputs than traditional edge detection or learned sketch generation; and (2) explicit design for enhanced user experience through expressive movements inspired by the ELEGNT framework, treating the robot not merely as a drawing tool but as a performing artist that communicates intentionality through choreographed gestures.

B. Image Processing for Line Art Generation

Traditional approaches to lineart extraction rely on edge detection algorithms including Canny, Sobel, and Laplacian operators. While computationally efficient, these methods struggle with noise sensitivity, parameter tuning requirements, and inability to distinguish foreground from background elements. Recent work has explored deep learning approaches including U-Net architectures for sketch synthesis, but most require extensive training datasets specific to portrait sketching. Stable Diffusion, based on latent diffusion models (LDMs) [5], enables high-fidelity image-to-image generation through a compressed latent space representation. The ControlNet architecture introduced conditional control for diffusion models[6], enabling precise image-to-image translation tasks. ControlNet-Lineart, specifically pretrained for line drawing extraction, provides robust performance across diverse input conditions without requiring task-specific fine-tuning. Our work applied this architecture to robotic drawing systems, demonstrating practical advantages for collaborative robot applications.

C. Expressive Movement in Robotics and Performing Arts

The ELEGNT framework [1] introduced a dual-utility model for robot movement design, combining functional objectives (task completion, efficiency) with expressive objectives (intention, attention, emotion). The framework proposes a Markov Decision Process (MDP) formulation where the total utility U combines functional utility U_f and expressive utility U_e:

$$U = U_f + \gamma \cdot U_e \tag{1}$$

where γ controls the balance between functional and expressive components. User studies demonstrated that expressive movements ($\gamma > 0$) significantly enhance engagement, particularly in social-oriented tasks, while maintaining task completion quality. However, ELEGNT introduced lamp-like robots with continuous presence as a prototype and did not address single-interaction creative tasks like portrait drawing. Beyond robotics, the concept of expressive gestures has deep roots in performing arts. The flourish—a decorative, exaggerated gesture used to draw audience attention or gracefully conclude a performance—appears across disciplines: orchestral conductors' final hand sweeps that signal musical completion while conveying interpretive emotion [7], stage magicians' presentation gestures (the "ta-da!" reveal) that transform functional object display into theatrical moments [8], and dancers' curtain call bows that acknowledge applause while expressing gratitude. These theatrical movements serve dual purposes: functional completion (signaling the end) and expressive communication (conveying pride, satisfaction, invitation to applaud). This connection between functional completion and expressive communication is well established in performing arts literature, where gestures serve both semantic and aesthetic purposes [7]–[9]. Such insights directly inform our design of the robotic flourish gesture as a performative act rather than a mechanical transition. Prior HRI studies demonstrate that subtle variations in robot body, head, or arm motion significantly influence user perception of communicative intent and affect [10], [11]. These findings motivate our use of expressive gestures not merely as aesthetic embellishment but as meaningful communicative cues within the interaction. Our work translates this performing arts principle into collaborative robotics by extending the ELEGNT framework to discrete creative interactions. This enables non-anthropomorphic robots to convey artistic intentionality through purposeful flourishes.

III. System Architecture and Implementation

Fig. 2. FIGUR portrait generation pipeline

Figure 2 illustrates the complete FIGUR system pipeline, consisting of five integrated modules: (1) intelligent acquisition, (2) generative stylization, (3) vectorization and path optimization, (4) robot coordinate transformation, and (5) expressive drawing execution with theatrical flourish gestures.

A. Intelligent Acquisition

Unlike conventional webcam capture that relies on manual user positioning, FIGUR implements an automated frontal-gaze detection system using MediaPipe Face Mesh. The system captures 300 frames over a 5-second interval while continuously computing a frontal-gaze score F for each frame:

$$F = \left(1 - |d_{x,\text{eyes}}|\right) \cdot \left(1 - |d_{y,\text{eyes}}|\right) \cdot \left(1 - |\theta_{\text{head}}|\right) \cdot S_{\text{symmetry}} \quad (2)$$

where $d_{x,\text{eyes}}$ and $d_{y,\text{eyes}}$ measure horizontal and vertical eye-alignment deviations from center, θ_{head} represents head rotation from frontal orientation, and S_{symmetry} quantifies bilateral facial-landmark symmetry. The system automatically selects the top-3 highest-scoring frames for subsequent processing, ensuring optimal input quality regardless of user experience level.

B. MLLM-Driven Stylization

FIGUR incorporates Google Gemini-2.5-flash to perform image-to-image translation. Unlike pixel-level filters, the MLLM interprets the semantic content of the input. We utilize prompt engineering to enforce strict rendering constraints: "pure black line-art," "white background," and "absence of shading." The system accepts user-defined style parameters (e.g., 'Ghibli', 'Webtoon') via the interface, modulating the prompt to alter the character design and line simplification level while preserving the subject's identity.

C. Vectorization and Path Optimization

Generated lineart images must be converted to vector paths suitable for robot pen control. We implement a two-stage pipeline integrating morphological skeletonization with greedy path optimization to ensure both aesthetic fidelity and kinematic efficiency.

1) Stage 1: Morphological Vectorization Pipeline: Traditional contour tracing algorithms (e.g., Potrace) typically extract the boundaries of a stroke, resulting in double-line artifacts that are unsuitable for thin-pen plotting. To resolve this, we employ the Zhang-Suen thinning algorithm [12] to extract the topological medial axis of the line art. The pipeline proceeds as follows: (1) The raster image undergoes Otsu's thresholding and morphological closing to repair minor discontinuities. (2) The binary foreground is skeletonized to a single-pixel width, preserving the structural connectivity of the subject. (3) Connected components are converted into vector polylines using chain approximation. To mitigate robotic jitter caused by excessive control points, we apply the Douglas-Peucker algorithm to approximate the curves, retaining essential geometric features while reducing data density.

2) Stage 2: Greedy Trajectory Optimization: The raw vector extraction yields N unordered strokes. Naïve sequential execution results in excessive non-drawing (pen-up) travel time and erratic manipulator movements. We formulate the path planning as a variant of the Traveling Salesperson Problem (TSP) and implement a greedy nearest-neighbor heuristic. Let P_{curr} be the current position of the end-effector. The algorithm iteratively selects the next unvisited stroke S_i that minimizes the transition distance from P_{curr} to either the start (P^i_{start}) or end (P^i_{end}) point of S_i. Crucially, the algorithm permits bidirectional traversal; if the endpoint is closer, the stroke vector is reversed. This optimization reduces the total air-time travel distance, directly contributing to the 2.4-minute average completion time.

D. Robot Coordinate Transformation

Vector paths defined in normalized image coordinates $[0, 1] \times [0, 1]$ must be transformed to the HCR-5 robot's base frame through a multi-stage transformation pipeline. First, we map normalized coordinates to physical paper dimensions while preserving aspect ratio and applying margins. Then, the paper origin $\mathbf{O}_{\text{paper}} = (x_0, y_0, z_0)$ is defined in robot base coordinates, with paper orientation θ relative to the robot X-axis. For each paper point (x_p, y_p), the robot coordinates are computed via a rotation transformation.

Each stroke generates a waypoint sequence: (1) move to stroke start with pen up (z = z_{travel}), (2) lower pen (z = z_{draw}), (3) traverse stroke points with pen down, (4) raise pen. Waypoints are serialized to JSON format with x, y, z, pen attributes for XML-RPC transmission to the robot controller.

E. Expressive Drawing Execution with Theatrical Flourish

Upon completion of the drawing, FIGUR performs an expressive presentation gesture inspired by performing arts traditions, rather than immediately retracting to home position (purely functional, $U_e = 0$). The robot executes a sweeping circular flourish that mimics the graceful hand movement of an artist stepping back to present their completed work. This choreographed gesture consists of three integrated phases:

1) Elevation Phase: Move to position above drawing center $P_{center} = (x_c, y_c, z_{observe})$ at observation height $z_{observe} = z_{draw} + 50$mm, with the pen-holding end-effector oriented toward the portrait. This establishes a "viewpoint" for admiring the completed work.

2) Circular Flourish Sweep: Execute a smooth, sweeping circular gesture in the XY-plane with radius r = 100mm, parameterized as:

$$x(t) = x_c + r \cdot \cos(2\pi t), y(t) = y_c + r \cdot \sin(2\pi t), t \in [0, 1]$$

The circular path is discretized into 20 uniformly distributed waypoints to ensure fluid motion. Velocity is reduced to 60% of the drawing speed ($v_{\text{express}} = 90\,\text{mm/s}$ vs. $v_{\text{draw}} = 150\,\text{mm/s}$), creating a deliberately theatrical pace that signals presentation.

3) Graceful Retreat with Vertical Ascent: During the circular sweep, the robot gradually elevates in the Z-axis: $z(t) = z_{observe} + (z_{home} - z_{observe}) \cdot t$ This creates a three-dimensional spiral trajectory, evoking a performer's courtly bow or a conductor's final flourish through combined XY-circular and Z-linear motion.

This gesture embodies expressive utility $U_e > 0$ in the ELEGNT framework, with velocity scaling that embodies the expressive utility parameter $\gamma = 0.6$. The circular path and

reduced speed convey organic intentionality, transforming the robot from a task executor to a collaborative artist. Post-study interviews (N=32) revealed that participants perceived the transition as shifting from "finishing the artwork" (functional) to "presenting the result with pride" (expressive), validating the design's impact on user engagement and perceived creativity.

IV. Experimental Results and Evaluation

We conducted comprehensive experiments to evaluate FIGUR's performance across three dimensions: (1) lineart generation quality, (2) vectorization and drawing accuracy, and (3) theatrical flourish impact on user experience.

A. Lineart Generation Quality Comparison

We evaluated the performance of our proposed Gemini-2.5-flash based pipeline against two baseline methods: (1) Canny Edge Detection[13] (traditional gradient-based method) and (2) DexiNed (Deep Learning-based Dense Extreme Inception Network for Edge Detection)[14]. The dataset consisted of 50 frontal face portraits captured via our acquisition system under varying lighting conditions. We assessed quality using three metrics: Line Continuity (percentage of stroke segments $> 10px$ without fragmentation), Noise Ratio (ratio of isolated artifact pixels to structural edge pixels), and Semantic Coherence (qualitative assessment of facial feature preservation).

TABLE I
Comparison of Line Art Vectorization Methods

Method	Line Continuity (↑)	Noise Ratio (↓)	Semantic Coherence
Canny + Otsu	62.4%	18.5%	Low
DexiNed	78.1%	8.2%	Medium
Proposed (Gemini 2.5)	94.3%	1.2%	High

As shown in Table I, our MLLM-driven approach significantly outperforms pixel-level baselines. While DexiNed captures detailed edges, it often produces excessive texture noise unsuitable for single-stroke robotic drawing. In contrast, the Gemini-2.5-flash pipeline achieves a 94.3% line continuity, demonstrating superior capability in abstracting semantic features into clean, continuous vectors. This reduction in fragmentation is critical for minimizing robotic pen-lift frequency.

B. Vectorization and Drawing Performance

We evaluated the efficiency of our Morphological Skeletonization and Greedy Optimization pipeline compared to a naive contour-following baseline. Metrics included Pen Travel Distance (non-drawing movement in air) and Total Drawing Time. The naive approach, which processes contours in the raw order of extraction, resulted in excessive cross-canvas transitions. Our proposed method, which employs skeletonization for centerline extraction combined with a greedy nearest-neighbor heuristic, achieved a 31% reduction in pen travel distance. This optimization directly translates to a reduced average completion time (2.4 minutes vs. 3.1 minutes for the baseline). Furthermore, the skeletonization process ensures consistent single-pixel stroke widths, improving the aesthetic quality of the physical output compared to variable-width contour filling.

C. Theatrical Flourish Impact: User Study

Qualitative Feedback:

Participants provided open-ended comments that explicitly referenced performing arts metaphors:

- "The sweeping flourish—like a magician's 'ta-da!' gesture—made it feel like the robot was proudly presenting its creation, not just finishing a task." (P18)
- "That elegant hand movement at the end, like a conductor finishing a symphony, completely changed how I saw the robot—from a machine to an artist." (P24)
- "Without the flourish, it just felt mechanical. The circular gesture with that slower pace made me actually want to applaud, like at the end of a performance." (P7)
- "The way it stepped back and presented the drawing reminded me of an artist unveiling a painting at a gallery—there was pride and intentionality in that movement." (P31)

These comments validate our design intent to translate performing arts gesture vocabulary into robotic movement, demonstrating that users readily interpret the theatrical flourish through familiar cultural references (conductors, magicians, gallery unveilings).

V. Conclusion

This paper presented FIGUR, an autonomous system that bridges the gap between semantic image understanding and physical robotic rendering. By integrating MLLM for stylistic abstraction and employing rigorous morphological processing and path optimization, FIGUR achieves a high level of artistic fidelity and mechanical efficiency. Future work will investigate the integration of closed-loop visual feedback, enabling the robot to self-correct drawing deviations in real-time.

Acknowledgment

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the National Program for Excellence in Software, supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP), in 2025.

References

[1] Y. Hu, P. Huang, M. Sivapurapu, and J. Zhang, "ELEGNT: Expressive and functional movement design for non-anthropomorphic robot," *arXiv preprint arXiv:2501.12493*, 2025.

[2] P. Schaldenbrand et al., "FRIDA: A collaborative robot painter with a differentiable, real2sim2real planning environment," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2023, pp. 11624–11630.

[3] G. Lee et al., "From scratch to sketch: Deep decoupled hierarchical reinforcement learning for robotic sketching," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2023, pp. 5602–5608.

[4] Y. Song, J. Park, M. Kim, and S. Lee, "Robot portrait drawing: A multi-stage approach to autonomous artistic rendering," *IEEE Trans. Robotics*, vol. 38, no. 4, pp. 2415–2428, 2022.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.

[6] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 3836–3847.

[7] K. Davidson, "The conductor's gesture: Meaning and effect in musical performance," *J. Musicology*, vol. 29, no. 3, pp. 275–297, 2012.
[8] T. Green, "Gesture in magic performance: The semiotics of the flourish," *Performance Research*, vol. 18, no. 5, pp. 89–98, 2013.
[9] F. Heider and M. Simmel, "An experimental study of apparent behavior," *American J. Psychology*, vol. 57, no. 2, pp. 243–259, 1944.
[10] S. Saerbeck and C. Bartneck, "Perception of affect elicited by robot motion," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2010, pp. 53–60.
[11] M. Hoffman, J. F. C. De Santis, and L. Wang, "Body, head, and eye movements during perception of directed communication from a humanoid robot," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2019, pp. 267–276.
[12] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," Commun. ACM, vol. 27, no. 3, pp. 236–239, 1984.
[13] J. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[14] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2020, pp. 1912–1921.