Panel-by-Panel Souls: A Performative Workflow for Expressive Faces in AI-Assisted Manga Creation

Qing Zhang
The University of Tokyo
qzkiyoshi@gmail.com,

Jing Huang
Tokyo University of the Arts
hkoukenj@gmail.com,

Yifei Huang
The University of Tokyo
hyf015@gmail.com,

Jun Rekimoto
The University of Tokyo
SONY CSL Kyoto
rekimoto@acm.org

Abstract

Current text-to-image models struggle to render the nuanced facial expressions required for compelling manga narratives, largely due to the ambiguity of language itself. To bridge this gap, we introduce an interactive system built on a novel, dual-hybrid pipeline. The first stage combines landmark-based auto-detection with a manual framing tool for robust, artist-centric face preparation. The second stage maps expressions using the LivePortrait engine, blending intuitive performative input from video for fine-grained control. Our case study analysis suggests that this integrated workflow can streamline the creative process and effectively translate narrative intent into visual expression. This work presents a practical model for human-AI co-creation, offering artists a more direct and intuitive means of "infusing souls" into their characters. Our primary contribution is not a new generative model, but a novel, interactive workflow that bridges the gap between artistic intent and AI execution.

1 Introduction

In manga (Japanese-style comics known for their expressive art and panel-based storytelling) a character's soul is conveyed in the subtle shift of an eye or the slight curve of a lip, a nuance that current text-to-image models consistently fail to capture. While state-of-the-art text-to-image models can generate aesthetically striking manga-style characters, they struggle with rendering the subtle, coordinated facial dynamics essential for sequential storytelling [3]. A manga artist's intent behind a "knowing glance," "a flicker of suspicion," or "a smile tinged with regret" is often lost in translation. This results in compositionally sound but emotionally vacant panels, or worse, expressions that are inconsistent from one panel to the next, breaking the narrative thread [3]. This "nuance gap" forces a manga artist into a laborious fine-tuning stage, manually redrawing faces to ensure the story flows correctly.

Existing digital solutions, however, present their own set of challenges for the practicing manga artist. Powerful facial reenactment tools are primarily designed for video production, with interaction models ill-suited for the iterative refinement of static manga panels [4, 5]. Alternatively, highly flexible toolkits like ComfyUI require artists to navigate complex, node-based editors. This disjointed process breaks the creative "flow state," which is crucial for a manga artist, who often works under tight deadlines and whose primary tool is a layer-based drawing application like Clip Studio Paint, not a visual programming environment. There is a clear gap for an integrated workflow designed

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

specifically for the narrative needs of a manga artist in terms of human-AI co-creation: one that is intuitive, direct, and allows for the fluid orchestration of expressions.

We address this need with a novel, dual-hybrid pipeline designed to balance generative AI efficiency with artistic control. Our workflow begins with a hybrid preparation stage that combines a landmark-based auto-detector for speed with a manual framing tool for handling complex hairstyles or accessories [7]. This is followed by a hybrid expression mapping stage, where broad emotions are captured through performative video input and then perfected with fine-grained numerical sliders [4, 7]. This multi-stage, artist-centric approach allows the manga artist to remain focused on character dynamics and emotional storytelling, using the system as a powerful, collaborative partner rather than a rigid tool.

Our primary contribution is a novel system and workflow that serves as a constructive model for human-AI co-creation. In the context of the changes and risks brought by generative AI, our work presents an alternative that leverages, rather than supplants, human creativity [8]. We offer both a concrete example of human-AI co-creation and a direct approach targeting the fundamental gap between text and nuanced expression. Specifically, this paper introduces a performative pipeline for managing individual character expressions within multi-character or sequential panels. This system contributes a practical tool and a new workflow that empowers artists to infuse their characters with emotion in a more direct and intuitive manner.

2 Related Work

2.1 The Expressive Challenge in Manga: Beyond the Text Prompt

The advent of powerful diffusion models like Stable Diffusion and DALL-E has enabled the generation of high-quality, manga-style illustrations from simple text prompts [22]. While adept at capturing overall artistic style, they frequently fail to render the subtle emotional states or micro-expressions with the precision required for sequential storytelling [23]. This issue stems from a fundamental cognitive-linguistic gap: our rich mental imagery of a character's expression is inherently resistant to complete verbal encoding [17, 18, 26, 24, 25]. Researchers have noted this barrier in general AI image creation, where users find it difficult to articulate their desired facial expressions with adequate precision [13, 15].

This challenge is amplified in manga, where storytelling relies on a unique and complex visual lexicon. A trope like *tsundere*, for instance, requires a subtle blend of anger, affection, and embarrassment [9] that defies a simple text description. Similarly, conveying the silent, panel-to-panel shift in a team captain's expression from steadfast confidence to quiet concern is nearly impossible to orchestrate through iterative text prompts [10]. Relying on text alone often forces the manga artist to accept generic, stereotypical emotions that flatten the narrative depth.

2.2 The manga artist's Digital Toolkit: Current Workflows and Frictions

To overcome the limitations of generative models, a manga artist typically reverts to their established digital toolkit, dominated by layer-based software like Clip Studio Paint [6]. Within this environment, two primary methods for expression control exist: direct manual drawing and the use of 3D posing models. Manually redrawing each face panel-by-panel offers maximum control but is intensely time-consuming and creates a high risk of introducing subtle inconsistencies in a character's appearance. The alternative, using built-in 3D posing dolls, helps maintain consistency but introduces new frictions. These models are often criticized for their generic appearance and stiffness, requiring significant manual alteration to match a character's unique style and inject life into their expression [27, 28].

While advanced visual programming interfaces for diffusion models, such as ComfyUI, offer powerful fine-tuning capabilities via node-based editors, their interaction paradigm is fundamentally disconnected from the manga artist's core drawing process. Requiring a manga artist to switch from a digital canvas to a complex graph editor breaks the creative "flow state" crucial for narrative work. These tools, while powerful, are not designed with the workflow of a sequential artist in mind, creating a clear human-AI interaction gap for a more integrated and intuitive solution [30].

2.3 The Workflow Gap: Integrating Reenactment for Co-Creation

Facial reenactment methods [16, 14, 2] offer compelling alternatives, enabling direct "performative" input to capture artists' desired expressions. However, these methods introduce their own significant workflow gap. State-of-the-art reenactment models, much like the powerful node-based editors discussed in Section 2.2, are rarely designed for a manga artist's panel-based workflow. They are often standalone tools for video production or exist as components within complex, non-intuitive interfaces that, as previously noted, break the creative "flow state".

The true gap is therefore the lack of a system that leverages generative AI for what it excels at—rapidly drafting scenes and characters—while seamlessly integrating an intuitive, performative system that allows the artist to re-take control and "infuse souls" into their characters' expressions panel-by-panel. Our work addresses this gap by integrating a high-fidelity reenactment engine into a novel interaction model, designed to bridge the needs of the manga artist with the capabilities of modern AI.

3 Our Approach: The Panel-by-Panel Pipeline

Our approach transforms an initial AI-generated manga panel into a narrative illustration through a three-stage pipeline designed for a manga artist's workflow: (1) Automated Face Preparation, (2) Interactive Expression Mapping, and (3) Layered Composition and Refinement. This process is designed to be repeated seamlessly across multiple characters and panels, allowing for the consistent development of emotional storytelling. The central hypothesis of our work is that the hybrid nature of the Interactive Expression Mapping stage is the key to this workflow's effectiveness, a claim we explore through a focused analysis in Section 4.

Stage 1: A Hybrid Pipeline for Artist-Centric Face Preparation. The pipeline begins with a hybrid face preparation stage, designed to balance automated efficiency with artist-centric control. We recognize that while automated face detectors are powerful, the creative scope of manga—with its complex hairstyles, face-obscuring accessories, and varied character orientations—demands a system that ensures the artist always has the final say.

Our system therefore provides two integrated modes: (1) Landmark-Based Auto-Detection: The primary mode utilizes a state-of-the-art face analysis model from the insightface library [21]. Instead of relying on a simple bounding box, this method detects 106 facial landmarks (eyes, brows, nose, mouth, and facial contour). We then algorithmically construct a tight, padded bounding box directly from these landmarks. This technique is highly effective at producing well-composed crops of human-like faces while explicitly excluding non-facial elements like hands and torsos, a critical failure point of general object detectors. (2) Manual Framing Mode: For cases where the automated detection is insufficient or when the artist wishes to make a specific creative choice (e.g., including a particular accessory as part of the 'face'), the system provides a seamless manual mode. The artist is presented with an interactive window where they can directly draw a square frame onto the source image. This manual bounding box is then treated identically to one generated by the auto-detector.

This dual-mode pipeline provides the speed of automation for standard cases while preserving the ultimate creative control and robustness of manual intervention, guaranteeing that any character can be prepared for the reenactment stage.

Stage 2: Interactive Expression Mapping. Continuing the hybrid design philosophy from the preparation stage, our Interactive Expression Mapping workflow is also built on a dual-mode interaction model designed to offer both intuitive performance and precise control. Our system leverages the pre-trained LivePortrait [14] model for its balance of real-time efficiency and high-fidelity output. We integrate this engine into a novel, hybrid interaction workflow that offers both intuitive performance and precise control.

First, the manga artist provides a performative input, either through a live webcam feed or by uploading a pre-recorded video reference, a common practice for artists seeking to capture specific expressions. Our interface presents this video on an interactive timeline, allowing the artist to scrub through the performance and select the single keyframe that best captures the desired macro-expression, such as the peak of a smile or a specific head tilt.

Recognizing that a single performance frame may lack precision, the system then offers a secondary stage of fine-grained numerical control. After the base expression is mapped, the manga artist

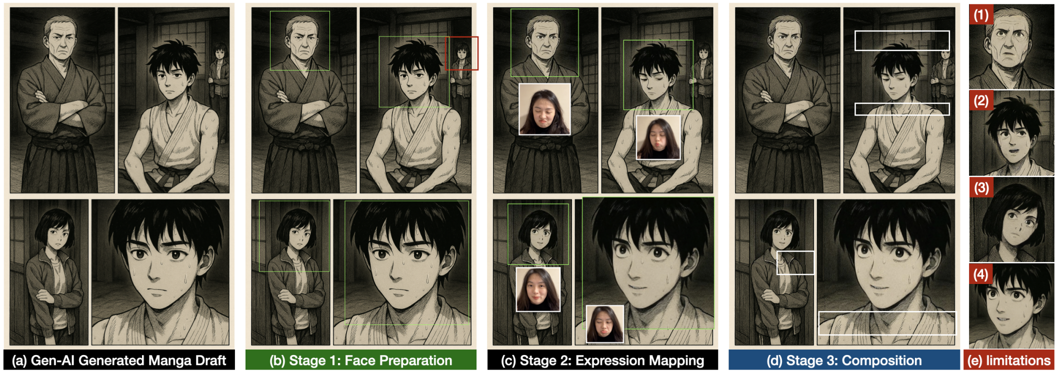
(a) Gen-AI Generated Manga Draft
(b) Stage 1: Face Preparation
(c) Stage 2: Expression Mapping
(d) Stage 3: Composition
(e) limitations
(1)
(2)
(3)
(4)

Figure 1: An end-to-end demonstration of our pipeline and an illustration of its current limitations. The process begins with (a) a Gen-AI generated manga draft with neutral character expressions. In (b) Stage 1: Face Preparation, our system successfully identifies and frames the primary faces, though it fails on a small, distant face (highlighted in red), illustrating a limitation of the auto-detector. During (c) Stage 2: Expression Mapping, new expressions are mapped onto the prepared faces from driving inputs (shown inset). In (d) Stage 3: Composition, the modified faces are re-integrated back into the draft, demonstrating a successful transfer of expression. Finally, (e) highlights typical limitations, which are largely inherited from the underlying LivePortrait model.

can use dedicated sliders to directly manipulate parameters of the eye and lip retargeting modules. This hybrid approach is allowing for adjustments that are difficult to perform, such as correcting a character's gaze independently of their head pose or precisely modifying the lip curvature to match a character's established design. This gives the manga artist the expressiveness of a reference performance combined with the meticulous control needed for detailed line art.

Stage 3: Composition and Refinement. The final stage of our pipeline focuses on the practical task of re-integrating the expressive face from Stage 2 back into the original manga panel. The process begins with a direct composition method: the 512x512 pixel reenacted face is first resized back to its original dimensions as recorded during Stage 1. It is then pasted onto the source panel at its precise original coordinates.

We acknowledge that this direct composition, while efficient, introduces predictable artifacts along the seams of the pasted region. These may include subtle geometric misalignments where the new face meets the character's neck and hair, or minor shifts in hue and lighting introduced by the reenactment model. A simple paste leaves a visible, jarring edge that requires refinement.

While automated seam-blending [1] could potentially address these artifacts, our work's primary contribution is the novel expression mapping pipeline (Stages 1 and 2). We therefore treat the final integration as a separate, known post-processing challenge. Our workflow intentionally concludes by handing off a composited draft, with its predictable artifacts, to the artist. This approach provides a clean hand-off point, allowing professional manga artists to apply their own expertise and familiar tools such as a smudge or airbrush brush in their preferred software for the final, nuanced polishing.

This design choice reinforces our vision of the system as a powerful assistant rather than a complete replacement. It automates the laborious task of redrawing faces for nuanced expressions—isolating characters and mapping complex expressions—while leaving the final, delicate task of aesthetic integration and polishing firmly in the hands of the artist. This respects their role as the ultimate arbiter of quality and maintains their creative control over the final artwork.

4 Analysis of an End-to-End Case Study

To evaluate the practical application and effectiveness of our pipeline, we conducted an end-to-end case study on a multi-character, multi-panel manga draft that was initially generated with neutral expressions using a state-of-the-art text-to-image model (DALL·E 3). The goal was to use our system to infuse the characters with new expressions to alter the narrative tone of the scene. The entire process and its outcomes are illustrated in Figure 1. This case study was conducted as a formative and expert evaluation. It was performed by the authors, who possess formal training in fine arts (visual communication design, painting) and human-computer interaction, allowing for a dual analysis of both the technical pipeline and its alignment with an artist's creative workflow.

Character Prompts: *"""Character 1 (Older Mentor): Middle-aged man with stern features, short graying hair, wearing a worn hakama and gi. Neutral, composed expression. Standing with arms folded, upright posture, authoritative presence. Character 2 (Young Trainee): Teenage boy, lean build, messy black hair, wearing a sleeveless gi. Sitting on the floor, legs crossed, looking forward with a blank, neutral expression. Sweat marks his brow, but his face is calm. Character 3 (Observing Peer): Teenage girl, short bob haircut, wearing a tracksuit top over training clothes. Leaning against the doorframe with one arm loosely folded. Neutral expression, gaze directed toward the seated boy."""*

Findings from Stage 1: Face Preparation. As shown in Figure 1(b), our landmark-based auto-detector successfully identified and framed the primary male characters. However, this process revealed two key insights. First, our findings suggest that characters with complex hairstyles or accessories often require the manual face extraction mode to achieve an artistically sound composition. This finding justifies our hybrid approach. The auto-detector's failure on the small, distant face (highlighted in red) further underscores the need for a manual override. Second, we found that a slightly generous crop that includes a character's hair works better for the subsequent reenactment stage than a tight crop on only the facial features. This provides more contextual information to the reenactment model, leading to a greater perceived coherency between the face and hairstyle in the final result.

Findings from Stage 2: Expression Mapping. Figure 1(c) illustrates the core expression mapping workflow. A crucial observation during this stage was the discrepancy between the driving performance and the final reenacted output. The most aesthetically pleasing or narratively correct facial expression on the manga character would often appear a few frames before or after the seemingly best frame of the driving video. This temporal offset highlights a key human AI interface challenge: the user's perception of their own best performance may not always align with the model's best interpretation. This finding underscores the necessity of our interactive timeline slider, which allows the artist to scrub through the results to find the optimal frame, rather than being locked into a single moment from their initial performance.

Findings from Stage 3: Composition and Limitations. The successful re-integration of the new faces is shown in Figure 1(d). However, the limitations of the underlying reenactment model become apparent here, as detailed in Figure 1(e). We observed that no matter whether we used relative or absolute motion modes in LivePortrait, a set of typical artifacts still appeared. These persistent artifacts include geometric inconsistencies where the hair and ears remain static during head rotation (e1-e3) and style mismatches where photorealistic features like lips and teeth are introduced into the monochrome aesthetic (e4). This observation reinforces that current reenactment methods are not a perfect one-shot solution.

5 Discussion

Our end-to-end case study provided several key insights into the practical challenges of human-AI co-creation for manga art. The analysis confirmed the necessity of our hybrid pipeline, as the automated face detector's failure on distant characters or complex hairstyles highlighted the need for a manual override to handle artistic diversity. More importantly, the expression mapping stage revealed a crucial human-AI co-creation challenge: a temporal offset, where the most aesthetically pleasing or narratively correct reenacted expression did not always align with the perceived best frame of the artist's driving performance.

This temporal discrepancy finding is central to our contribution. It demonstrates that a simple, one-shot "perform-and-generate" model is insufficient for this creative task. Instead, it validates our workflow's emphasis on an interactive timeline slider, which allows the artist to scrub through the results and discover the optimal frame. This transforms the interaction from a rigid command into a more fluid, collaborative exploration.

The importance of this human-in-the-loop approach is underscored by existing concerns within the professional manga community. In an expert interview conducted by the authors after the case study, a professional manga artist noted that the current widespread adoption of digital tools (e.g., pen tablets with line smoothing) has already led to a perceived “stylistic homogenization”—a loss of the unique, individual “pen stroke” characteristic of traditional, physical-media artwork.

This existing concern over stylistic homogenization provides a crucial parallel to the problem of expressive homogenization our paper addresses: the “emotionally vacant” or generic faces produced by text-to-image models. Our workflow, therefore, can be seen as a counter-movement. Instead of further automating and “smoothing” the artist’s unique hand, it is designed to capture and translate the artist’s personal, performative intent. It serves as a constructive model for human-AI co-creation by re-centering the artist, empowering them to “infuse souls” through their own performance while leaving the final aesthetic integration firmly in their control.

Despite its successes, our proof-of-concept also has clear limitations. The most significant is the lack of holistic, 3D-aware understanding, which results in artifacts where a character’s head rotates but their hair and ears remain static. Furthermore, the stability of the facial reenactment is sensitive to head pose. Echoing the challenges faced by foundational models [23], we observed that faces turned more than 45 degrees from the camera are still problematic, leading to less reliable expression mapping. Style mismatches can also occur, where photorealistic features are introduced into a stylized artwork, breaking the aesthetic cohesion and suggesting the necessity of a fine-tuned model specifically for this artistic domain. Finally, our current landmark-based auto-detector, while effective for human-like faces, is not designed to handle the full range of non-human characters common in manga creation. We acknowledge that facial reenactment technologies carry dual-use risks, but our work’s focus remains on the constructive application of these tools to empower artists.

These limitations provide a clear roadmap for future work. To address the geometric inconsistencies, the next iteration of our system should integrate 3D-aware models, potentially leveraging ControlNet-like mechanisms for pose control [29]. Such an approach would allow artists to manipulate not just the facial muscles but the character’s entire head orientation and posture, offering unparalleled freedom for narrative creation. To combat style mismatch, future research could explore fine-tuning the reenactment model on a curated dataset of manga-style art or incorporating style-preserving techniques.

Ultimately, we see our system as a step that could transform the artist’s role from a puppeteer of expressions to a true digital sculptor of character performance. By offloading the laborious task of redrawing faces for every subtle emotional shift, our work suggests that the most promising path for creative AI is not one that seeks to replace the artist, but one that provides them with powerful, intuitive, and collaborative tools to bring their visions to life more effectively. A key direction for future work is to conduct a formal user study with professional manga artists, moving beyond our initial author-led evaluation, to quantitatively assess our system’s impact on workflow efficiency and qualitative ratings of expressive control. The expert interview conducted for this paper confirmed the relevance of our research problem. The key human-AI co-creation challenges identified in our initial case study—such as the temporal offset in expression mapping and the necessity of the manual framing tool for complex compositions—will serve as the primary hypotheses and design probes for this future formal evaluation with industry professionals.

6 Conclusion

This paper introduced a dual-hybrid, performative workflow for AI-assisted manga creation. Our analysis indicates that this artist-centric approach is potentially a viable and effective method for translating narrative intent into visual art. This research contributes a practical pipeline that streamlines a tedious creative task, serving as a constructive model for human-AI co-creation in visual storytelling.

Acknowledgments and Disclosure of Funding

We sincerely thank professional manga artist YUNA (Mingxi Zhang) for her invaluable time and expert insights into the professional digital manga workflow. This work was supported by JST Moonshot R&D Grant JPMJMS2012, JSPS Grant-in-Aid for Early-Career Scientists JP25K21241, JSPS KAKENHI JP25K24384, and JPNP23025 NEDO.

References

[1] Zhang, L., Wen, T., & Shi, J. (2020). Deep image blending. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 231–240).

[2] Zhao, X., Xu, H., Song, G., Xie, Y., Zhang, C., Li, X., Luo, L., Suo, J., & Liu, Y. (2025). X-NeMo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*.

[3] Wu, J., Tang, C., Wang, J., Zeng, Y., Li, X., & Tong, Y. (2025). Diffsensei: Bridging multi-modal LLMs and diffusion models for customized manga generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 28684–28693).

[4] Luo, Y., Rong, Z., Wang, L., Zhang, L., Hu, T., & Zhu, Y. (2025). Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. *arXiv preprint arXiv:2504.01724*.

[5] Liao, L., Kang, L., Yue, T., Zhou, A., & Yang, M. (2025). Enhancing Facial Expressiveness in 3D Cartoon Animation Faces: Leveraging Advanced AI Models for Generative and Predictive Design. *International Journal of Advanced Computer Science & Applications*, 16(1).

[6] CELSYS, Inc. (2025). CLIP STUDIO PAINT — clipstudio.net. *https://www.clipstudio.net/*. [Accessed 28 October 2025].

[7] Yu, H., Mallick, R., Betke, M., & Bargal, S. A. (2025). GenEAva: Generating Cartoon Avatars with Fine-Grained Facial Expressions from Realistic Diffusion-based Faces. *arXiv preprint arXiv:2504.07945*.

[8] Mendoza, A. C. (2025). AI tools as companions in manga creation — Haribon Publishing — haribonpublishing.com. *Haribon Publishing*. Available at: https://www.haribonpublishing.com/blog/ai-tools-as-companions-in-manga-creation#google_vignette. [Accessed 10 August 2025].

[9] Abbott, M., & Forceville, C. (2011). Visual representation of emotion in manga: Loss of control is Loss of hands in Azumanga Daioh Volume 4. *Language and Literature*, 20(2), pp. 91-112.

[10] Johnson, D. (2024). "The Emotions that Get Stuck in Your Throat": Expressivity in Speech, Script, and Sound in Japanese Animation. *Animation Studies*. Available at: https://oldjournal.animationstudies.org/daniel-johnson-the-emotions-that-get-stuck-in-your-throat-expressivity-in-speech-script-and-sound-in-japanese-animation/. [Accessed 10-08-2025].

[11] Oshiba, J., Iwata, M., & Kise, K. (2023). Face image generation of anime characters using an advanced first order motion model with facial landmarks. *IEICE TRANSACTIONS on Information and Systems*, 106(1), pp. 22-30.

[12] Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 252, pp. 124260.

[13] Karras, T., Aittala, M., Aila, T., & Laine, S. (2023). Generative neural texture synthesis. *ACM Transactions on Graphics*, 42(1).

[14] Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., & Zhang, D. (2024). LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168*.

[15] Meléndez Fuentes, N. (2018). *Beyond Words: The Limits of Linguistic Expression. How to Express It All?*.

[16] Rochow, A., Schwarz, M., & Behnke, S. (2024). FSRT: Facial Scene Representation Transformer for Face Reenactment from Factorized Appearance Head-pose and Facial Expression Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7716–7726).

[17] Peper, A. (2022). A general theory of consciousness II: The language problem. *Communicative & Integrative Biology*, 15(1), 182–189.

[18] Schiffer, S. (2017). Intention and Convention in the Theory of Meaning. *A Companion to the Philosophy of Language*, 49–72.

[19] Yuan, Y., Zeng, J., & Shan, S. (2023). Describe Your Facial Expressions by Linking Image Encoders and Large Language Models. In *BMVC* (p. 377).

[20] Springbord. (2023). *Challenges Of Data Labelling And How To Overcome Them*. Available at: https://www.springbord.com/blog/challenges-of-data-labeling/ [Accessed: January 12, 2025].

[21] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690–4699).

[22] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., & Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6007–6017).

[23] Wu, Y., Xu, H., Tang, X., Chen, X., Tang, S., Zhang, Z., Li, C., & Jin, X. (2024). Portrait3D: Text-Guided High-Quality 3D Portrait Generation Using Pyramid Representation and GANs Prior. *arXiv preprint arXiv:2404.10394*.

[24] Monzel, M., Karneboge, J., & Reuter, M. (2024). Affective processing in aphantasia and potential overlaps with alexithymia: Mental imagery facilitates the recognition of emotions in oneself and others. *Biomarkers in Neuropsychiatry*, 11, 100106.

[25] Nakabayashi, K., & Mike Burton, A. (2008). The role of verbal processing at different stages of recognition memory for faces. *European Journal of Cognitive Psychology*, 20(3), 478–496.

[26] Mathews, A., Ridgeway, V., & Holmes, E. A. (2013). Feels like the real thing: Imagery is both more realistic and emotional than verbal thought. *Cognition & emotion*, 27(2), 217–229.

[27] Endangered AI. (2024). *Combine FaceID and Facial Expressions with IPAdapter & Controlnet*. Available at: https://www.youtube.com/watch?v=IlOOzmQZBzU [Accessed: January 13, 2025].

[28] Aiconomist. (2024). *How to Control ANY Facial Expression - ComfyUI Advanced LivePortrait Tutorial*. Available at: https://www.youtube.com/watch?v=OzAHpnCa_sc [Accessed: January 13, 2025].

[29] Zhang, L., Rao, A., Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3836–3847).

[30] Not4Talent. (2024). *Full FACIAL EXPRESSION control for Stable Diffusion (+Lora Pack)*. Available at: https://www.youtube.com/watch?v=Lg2WeL5lVFY [Accessed: January 13, 2025].

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim the contribution of a novel, dual-hybrid pipeline for AI-assisted manga expression. These claims are substantiated in Section 3, which details the system's architecture, and Section 4, which provides a case study and workflow analysis validating the design.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Section 5 ("Discussion") contains a detailed discussion of the work's limitations, including artifacts from the 2D reenactment model (e.g., static hair), sensitivity to extreme head poses, potential style mismatches, and the auto-detector's limitations with non-human characters.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper presents a novel interactive system and a Human-AI-Interface-focused workflow analysis. It is an empirical work and does not contain formal theoretical results, theorems, or mathematical proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 ("Our Approach") details the three-stage pipeline, naming the core pre-trained models used (InsightFace, LivePortrait). Section 4 ("Analysis") describes the case study setup and provides the prompts used to generate the source material, offering sufficient detail for conceptual reproduction of the workflow.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: To facilitate full reproducibility, we have made our code available in a repository. The assets can be found at: `https://github.com/artisticsciencex/manga_face_reenact`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the pre-trained models used for each stage of the pipeline (e.g., InsightFace for detection, LivePortrait for reenactment) in Section 3. Section 4 provides the prompts used to generate the source material for the case study. As the work uses pre-trained models and a qualitative case study, there are no training hyperparameters to report.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper's evaluation is based on a qualitative, end-to-end case study (Section 4). The evaluation criteria are the successful application of the workflow and the quality of the narrative output, not quantitative metrics for which statistical significance tests would be appropriate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The pipeline was run on a consumer-grade desktop computer equipped with an i9 12900 CPU with 128GB RAM, and a NVIDIA ada A6000 GPU. The models used (InsightFace, LivePortrait) are well-established and known to run effectively on modern consumer GPUs or powerful CPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presents a creative AI tool for artists. The work respects intellectual contributions by citing the models used, and its focus on human-AI collaboration aligns with the ethical goal of augmenting, rather than replacing, human creativity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper's Discussion section focuses on the potential positive impacts for artists. It also acknowledges the risks of generative AI and that facial reenactment technology can be misused, though the work's focus is on constructive, artistic applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The research uses existing, publicly available models (InsightFace, LivePortrait) and does not release a new, high-risk generative model or dataset. Therefore, specific safeguards for a new asset release are not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the primary models used (InsightFace, LivePortrait) and provides citations to their respective papers and repositories in Section 3. The licenses for these assets are respected in our proof-of-concept implementation.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, the proof-of-concept code is provided as a new asset. Documentation to ensure reproducibility is provided in the repository, as noted in our answer to Question 5.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper's primary case study was performed by the authors. This was supplemented by a formative, informal interview with one professional manga artist to contextualize our problem, as described in Section 5.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The performative aspect of our workflow involving the authors... was reviewed and approved by our institution's Institutional Review Board (IRB). The supplementary expert interview (Section 5) was conducted with verbal consent and did not involve sensitive data, falling under our institution's exempt research guidelines.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The core methodology of this research is based on computer vision models for face analysis (InsightFace) and reenactment (LivePortrait). Large Language Models were not a component of the system's architecture or the experimental workflow.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.