CraftGraffiti: Exploring Human Identity with Custom Graffiti Art via Facial-Preserving Diffusion Models

Ayan Banerjee, Fernando Vilariño, Josep Lladós
Computer Vision Center, Universitat Autònoma de Barcelona
{abanerjee, fernando, josep}@cvc.uab.cat

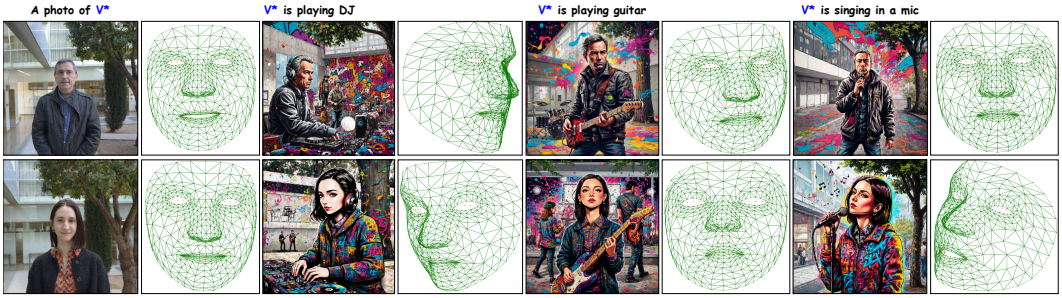

Figure 1: **Customization with *CraftGraffiti*:** It can accurately customize the input character pose in three different musical scenarios, preserving the facial attributes in graffiti style, as the only the pose of the face mesh changes with style while preserving the mesh structure of the input image.

Abstract

Preserving facial identity under extreme stylistic transformation remains a major challenge in generative art. In graffiti, a highly contrasted abstract medium, subtle distortions to the eyes, nose, or mouth can erase the recognizability of the subject, undermining both the personal and cultural authenticity. We present *CraftGraffiti*, an end-to-end text-guided graffiti generation framework designed with facial feature preservation as a primary objective. Given an input image and a style and pose descriptive prompt, *CraftGraffiti* first applies graffiti style transfer via LoRA-fine-tuned pretrained diffusion transformer, then enforces identity fidelity through a face-consistent self-attention mechanism that augments attention layers with explicit identity embeddings. Pose customization is achieved without keypoints, using CLIP-guided prompt extension to enable dynamic re-posing while retaining facial coherence. We formally justify and empirically validate the "style-first, identity-after" paradigm, showing it reduces attribute drift compared to the reverse order. Quantitative results demonstrate competitive facial feature consistency and state-of-the-art aesthetic and human preference scores, while qualitative analyses and a live deployment at the Cruïlla Festival highlight the system's real-world creative impact. *CraftGraffiti* advances the goal of identity-respectful AI-assisted artistry, offering a principled approach for blending stylistic freedom with recognizability in creative AI applications. The code is available at: github.com/ayanban011/CraftGraffiti.

1 Introduction

Human identity preservation in the GenAI era - bias, aesthetics, and open experimentation: Generative systems such as GANs and diffusion models [3, 29] replicate and amplify cultural biases from their training data, entering into the risk of distorting human identity across multiple dimensions [15]. Since usually trained on not particularly curated datasets, it is known that current generative

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

systems usually tend to disproportionately produce light-skinned, youthful, conventionally attractive faces, marginalizing other demographics, and concentrating on a narrow band of a specific concept of aesthetics [33]. Additionally, both children and older adults are usually either underrepresented or not properly identified, even for age-progression models [33, 30], presenting results that tend to transform the image of a young person into an adult face, or conversely, elderly people as unrealistically younger figures. Errors in the gender identification are usually more prone to affect women than men, and reinforce gender norms and heteronormativity, as seen in occupational and presentational stereotypes [45, 59]. Rooted in imbalanced datasets, these distortions risk homogenizing outputs unless addressed through curated data, fairness-aware training, and interdisciplinary critique [13].

For this reason, and in light of the relevant risks identified, it is essential to openly address the tensions to which our human identity is exposed in the new digital era; an era of interconnected individuals endowed with generative capacities. Such a topic is profound and complex from the philosophical, ethical, political, and technical dimensions [5, 2]. We state that in order to address the discussion and public debate of these topics, we can make use of two main instruments: 1) An artistic installation providing a representation of the customized human pose, technically sound and preserving facial attributes as much as possible, and 2) an ecological environment that allows for the natural interaction with people, providing the explicit context of a living lab or open experimentation space.

Integrating Cultural Expression and Facial Consistency in Generative Street Art: For the first aim, we need a cultural bind that serves as an entry point for people to interact with this new era of GenAI, which has to be inclusive and accessible for all; from this perspective, graffiti appears as an excellent candidate. Graffiti is a culturally significant art form, merging creative expression with the diversity of human identity [10, 14] and often serving as a voice for marginalized communities [4]. While recent advances in GANs [54, 55] and diffusion models [57, 51] offer automated generation, existing methods struggle to preserve facial attributes across varying styles and poses. Text-based image editing [24, 21] and style transfer models [20, 51, 17] fail to simultaneously adjust poses, while pose editing approaches [8, 36] lack unified style transfer. Unified models [38, 31, 11] attempt both but often lose facial details due to self-attention limitations.

So we introduce *CraftGraffiti*, an end-to-end generative framework that produces custom graffiti-style portraits of human characters while explicitly preserving the subject's facial identity. From the face mesh analysis of Fig. 1 it can be concluded that, *CraftGraffiti* preserves the facial structure as we always obtain same face mesh in different pose. This system is text-guided, allowing a user to specify high-level context (e.g., prompting a portrait of the person as a DJ, guitarist, or singer) and to customize the character's pose, all within a graffiti art style. To accomplish this, *CraftGraffiti* integrates several components: a CLIP-based text encoder for semantic guidance, a graffiti-style fusion module implemented via a pre-trained diffusion transformer [1] fine-tuned on graffiti aesthetics, a LoRA-based adapter [19] that efficiently injects pose information without requiring full model retraining, and a novel face-consistent self-attention mechanism that ensures key facial features of the input are maintained throughout the diffusion process. The final output is rendered by a variational autoencoder (VAE) decoder, yielding a high-quality graffiti portrait that retains the individual's identity enabling simultaneous pose adjustment.

Our contributions include: (1) a **unified framework** for pose-guided, graffiti-themed portrait generation that preserves individual identity via **mitigating gender bias**; (2) a novel **self-attention module** specifically designed for **facial consistency** in diffusion-based generation; and (3) state-of-the-art performance in both qualitative visual fidelity and quantitative identity-preservation metrics compared to existing methods.

A public artistic installation in an ecological environment as a living lab: For our second aim, we propose to explore ecological environments to enhance user-experience experimentation by enabling real-world engagement in authentic contexts [26]. Living labs provide user-centred, open innovation ecosystems with real-world experimentation, co-creation, and multi-stakeholder collaboration [9, 1]. Outdoor labs with pervasive sensing and crowdsensing further increase scalability and context-awareness [41], fostering inclusivity, diversity, and richer insights than artificial environments. From this perspective, music festivals appear as excellent ecological experimentation spaces for this type of trans- and multi-disciplinary investigation. In our case, we set our installation during the four days of the Festival Cruïlla, taking place 9-12 July 2025 (Please refer to the appendix for more details of the installation setup, outcomes, and human feedback).

[1] https://civitai.com/models/1058970/graffiti-style-fluxFaces

2 CraftGraffiti

Our proposed methodology consists of three major components: (1) text-to-image diffusion models, (2) Style Fusion mechanism, and (3) text-guided pose customization. A detailed discussion of these components has been obtained in the appendix.

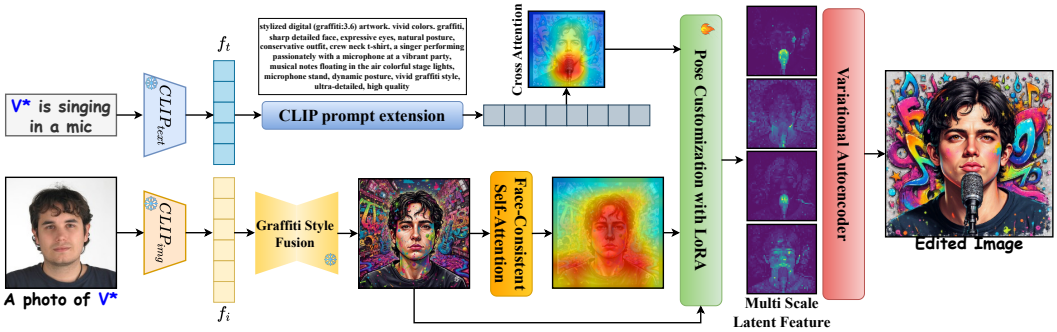

Figure 2: ***CraftGraffiti*** transforms a source image into a graffiti-style portrait while preserving the subject's identity and pose. Graffiti style is injected via a pretrained diffusion fine-tuned with LoRA for the dedicated style. Later on, another diffusion model is equipped with face-consistent self-attention and cross-attention modules to preserve key facial features, and a LoRA module enables pose customization without full model retraining via CLIP-based prompt extension. Finally, multi-scale latent feature processing using a VAE ensures that both global structure and fine details are captured across different resolutions in the latent space, yielding a high-quality graffiti-style image.

2.1 Why should we perform graffiti style transfer before facial attribute preservation

Performing the style transform before enforcing attribute constraints can be seen via a simple model. Let I be the original face image and let $S(I)$ be its graffiti version under a LoRA-fine-tuned diffusion model. If we have a facial-attribute extractor $F_a(I) \in \mathbb{R}^k$ (encoding eyes, nose, chin, expression, etc.) and an attribute-loss $\mathcal{L}_{attr}(X) = ||F_a(X) - F_a(I)||^2$, we can choose a projection $P(\cdot)$ so that for any image X, $F_a(P(X)) = F_a(I)$ (i.e. P restores I's attributes). Then by reconstruction $L_{\text{attr}}(P(S(I))) = 0$, whereas $P(I) = I$ (since I already has its own attributes), so $\mathcal{L}(S(P(I))) = \mathcal{L}(S(I)) = ||F_a(S(I)) - F_a(I)||^2 > 0$ whenever S perturbs the attributes. In other words, $\mathcal{L}_{attr}(P \circ S(I)) \leq \mathcal{L}_{attr}(S \circ P(I))$. This shows formally that applying the style S first and then the attribute correction P yields lower attribute discrepancy. This matches intuition from prior work: style-transfer methods aim to alter appearance while preserving its underlying structure [42]; however, if the style influence is too strong, it undermines the structural integrity of the content [47]. Moreover, LoRA fine-tuned diffusion models often use the early timesteps to reconstruct coarse content and later steps to add stylistic detail [37]. Thus, we preserve identity better by first adding the graffiti style and finally preserving the facial features, rather than applying constraints before styling. The overall methodology of *CraftGraffiti* has been demonstrated in Fig. 2.

Theorem. Let I be the input face with attributes $a = F_a(I)$. Let S be a diffusion-style operator and P an operator satisfying $F_a(P(X)) = a$ for all X (so P restores I's attributes). Then the image $X' = P(S(I))$ obeys $F_a(X') = a$ (it exactly preserves the original attributes), whereas $X'' = S(P(I))$ satisfies $F_a(X'') = F_a(S(I))$ and in general $F_a(S(I)) \neq a$ unless S itself preserved F_a. Hence $(P \circ S)(I)$ maintains the facial attributes exactly, whereas $(S \circ P)(I)$ need not.

Proof. By assumption $F_a(P(X)) = a = F_a(I)$ for any input X. In particular, $F_a(P(S(I))) = a$, so $P(S(I))$ has the original attribute vector. On the other hand, since $P(I) = I$ (the identity image already has attributes a), we have $S(P(I)) = S(I)$ and hence $F_a(S(P(I))) = F_a(S(I))$. Unless S is attribute-preserving by itself, $F_a(S(I)) \neq F_a(I)$ in general. Therefore, $P(S(I))$ preserves a exactly, while $S(P(I))$ typically does not, completing the proof.

2.2 How the extra dimension in the self-attention helps in facial attribute preservation

In Diffusion models [60, 50], each self-attention layer computes attention only among features of the current latent (no cross-latent computation). Concretely, the latent feature map of size $H \times W$

is reshaped into a sequence of tokens, and for each token, a query (Q), key (K), and value (V) vector are computed from the same latent features. The attention weights are then obtained by taking dot products of Q and K across all positions (scaled by affinity distance $\sqrt{d}$) and passing through softmax, and the result reweights the V vectors (see Fig. 3). This means every pixel in the latent try to see every other pixel in the same latent when computing self-attention.

To ensure that the two generated characters depict the same identity, we introduce an explicit identity embedding into the attention computation. Conceptually, this means adding a special identity vector (often derived from a reference face) as an extra token or feature dimension in the self-attention layer. In practice, one can concatenate an identity embedding to every spatial token (equivalently, add an extra "CLS"-style token) or append an identity channel to the latent features. The result is that queries and keys now carry identity information as well as image content. In effect, each pixel's query includes a fixed identity code, and each key (or value) can be augmented by the same code, biasing attention towards features matching that identity.

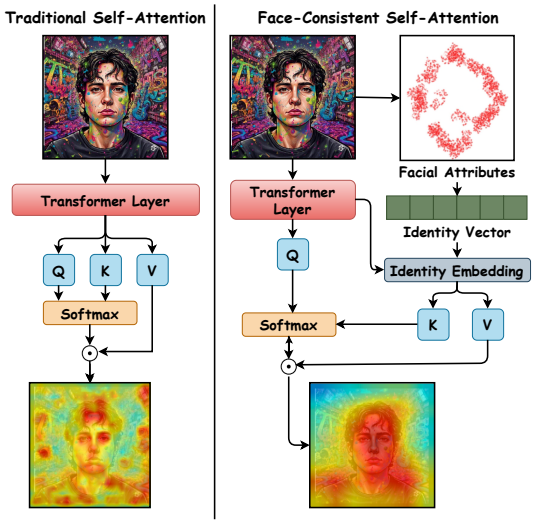

Figure 3: **face-consistent self-attention:** we can easily preserve the facial attribute of the character through the extra dimension of identity embedding.

In an identity-enhanced self-attention, the attention matrix is computed using both the spatial features and the identity embedding. For example, one might form $Q' = W_Q[x; id]$, $K' = W_K[x; id]$ where x is a spatial feature and id is the identity vector. Then, attention weights come from $Q'K'^T$. This extra dimension (the identity channel or token) causes the softmax to favor aligning features that correspond to the same identity. Architecturally, the U-Net is unchanged except at attention layers: we simply extend the input feature dimension by the identity embedding. The key effect is that the same identity code is shared across the entire image and across images in a batch. As a result, when generating multiple images of the same person, the self-attention layers see the same identity embedding each time, forcing consistent facial features.

3 Experimental Results

Qualitative Analysis: We compare *CraftGraffiti* against several FLUX-based approaches, FLUX + IP-Adapter [56], and FLUX.1 Kontext [27], however, both approaches fail to generate the graffiti style. Although FLUX + IP-Adapter [56] tried to maintain facial consistency, it was unable to change style. On the other hand, FLUX.1 Kontext [27] tries to change the background into a more cartoonist style, and the faces are not consistent. This shows how hard to unify the style transfer and consistency tasks and justify our decision to perform style transfer before facial consistency (see Fig. 4).

We also compare *CraftGraffiti* with InstructPix2Pix [6]; however, it neither adds objects nor generates high-quality images. Similarly, VLMs (Grok 3 [16] and GPT4o [22]) maintain consistency, unable to blend style due to the complexity of graffiti style transfer. This empirically shows the complexity of the problem (facial consistency in graffiti style) we are tackling and demonstrates the correctness of the component choice of *CraftGraffitii* for accurate style blending by preserving the facial features (For more qualitative examples, please refer to the Fig. 10 of the appendix.)

Quantitative Analysis: Table 1 presents a quantitative comparison against the state-of-the-art image editing techniques. *CraftGraffiti* stands in the fourth position after GPT4 [22], GPT4o [22], and [56] respectively, in terms of facial feature consistency (FFC), as it is very difficult to compare the preservation of facial features, as most of the feature extraction models (InceptionResNetV1) consider global features rather than the local ones. However, during style transfer, it is hard to preserve the global features of the input images. As other image editing [16, 22, 56] models are unable to blend the graffiti style, preserve the global features, and perform better in this metric.

Figure 4: **Qualitative Comparison:** CraftGraffiti perfectly transforms the input image into the graffiti style and maintains facial attributes, while the rest cannot do both.

Table 1: Quantitative Evaluation of CraftGraffiti

Method/Metric	FFC ↑	Aes ↑	HPS ↑	Inf. Time (sec) ↓
FLUX + IP-Adapter [56]	0.8324	3.2414	0.3012	8.2
FLUX.1 Kontext [27]	0.6741	2.1749	0.2911	6.7
InstructPix2Pix [6]	0.7112	2.7272	0.1918	3.1
GPT4o [22]	**0.8761**	4.5193	0.3412	13.4
Grok 3 [16]	0.8513	4.1652	0.3102	11.2
Ours (Baseline)	0.7618	3.6913	0.2911	**2.9**
Ours (+ Style Fusion)	0.6814	4.7195	0.3001	5.3
Ours (+ Face Consistent self attention)	0.7713	5.1376	0.3176	8.7
CraftGraffiti	0.7713	**5.2271**	**0.3536**	10.1

On the other hand, Aesthetic Score [32] and Human Preference Score [53] measure how perfectly the models blend the graffiti style with the input images. *CraftGraffiti* outperforms all the state-of-the-art image editing techniques in both metrics, as they are unable to blend the graffiti style properly over the input images. Last but not least, *CraftGraffiti* has very little inference time compared to the VLMs [16, 22], which makes it suitable for real-time applications. Further ablation studies of *CraftGraffiti* are available in the Fig. 6 and Fig. 7 of the appendix.

Human Evaluation: We have conducted the perceptual study in order to evaluate the credibility and widespread adoption of the graffiti synthesized with *CraftGraffiti*. We chose 60 graffiti (20 for each pose) synthesized by [56, 27, 6, 16, 22] and *CraftGraffiti* and asked the users to rate the graffiti on a scale of 1 to 5 (1 Low; 5 high) based on aesthetics, style blending, and recognizability. 47 anonymous users, 11 countries, participated in this study. The outcome of this user study is reported in Fig. 5.

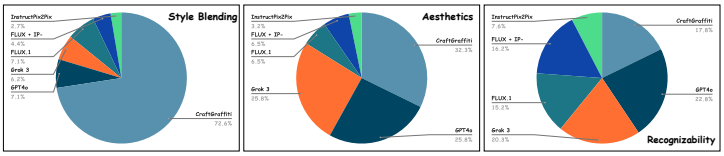

Figure 5: **Human Evaluation:** *CraftGraffiti* outperforms SOTA techniques [56, 27, 6, 16, 22] in style blending and aesthetics while preserving facial attributes (a decent performance in recognizability).

For the user, *CraftGraffiti* makes a perfect graffiti-style poster, while other models either change the background or make it more cartoonish. In terms of recognizability *CraftGraffiti* preserves the local features (eyes, nose, lips, and so on) but changes the global features during style transfer, maintaining a decent recognizability compared to the rest. The evaluators were also asked to rate cultural resonance. We have discussed it in the Section 4.

4 Discussion on *CraftGraffiti* Performance

Technical Contributions and Performance: Our results demonstrate that integrating a face-consistent self-attention mechanism with LoRA-based fine-tuning allows *CraftGraffiti* to preserve key facial attributes across diverse poses and graffiti styles. Compared to existing pose-editing and style-transfer methods, *CraftGraffiti* achieves state-of-the-art identity preservation without compromising artistic quality. This balance between recognizability and stylization is essential for applications where personal identity and cultural authenticity are intertwined.

Cultural and Societal Implications: Beyond technical performance, *CraftGraffiti* acts as an enabler in ongoing cultural debates around representation, bias, and access to AI-generated art. By using graffiti, a medium historically tied to marginalized voices, we highlight the potential for generative systems to serve as inclusive cultural tools. However, these systems also risk reproducing and amplifying harmful biases, which must be addressed through dataset curation, fairness-aware training, and critical engagement with affected communities. By deploying the installation at Festival Cruïlla, we had an opportunity to observe user interactions in a high-energy, socially diverse, real-world context. This aligns with the principles of ecological validity [26] and living lab methodologies [9, 1]. The festival setting facilitated spontaneous engagement, enabling us to capture authentic emotional responses and uncover interaction patterns that may not emerge in controlled lab settings [41]. Such environments also pose technical challenges in terms of managing environmental noise, variable lighting, unpredictable crowd dynamics, and regulatory uncertainties. In our specific case, it appeared as an excellent example to dry test the new European AI Act. The implications of this regulatory learning approach are of enormous interest, though they are beyond the scope of this piece of work.

Limitations: While *CraftGraffiti* successfully preserves facial attributes, it remains constrained by the biases present in its pretrained baselines, which may affect demographic and stylistic diversity. Also, our evaluation was conducted in a single cultural context (a European music festival), which may limit the generalizability of the findings to other socio-cultural settings. However, general considerations can be taken regarding the questions posed: *CraftGraffiti* produces graffiti-styled human figures that preserve the original subject's identity by maintaining key facial attributes across poses and styles through a face-consistent self-attention mechanism. Unlike many generative systems, it inherits body shape from the input image, avoiding homogenization into a single aesthetic norm. From an anthropological perspective, the tendency of many AI models to generate light-skinned, symmetrical, youthful, and slender bodies reflects the reinforcement of dominant, often Eurocentric, beauty standards [33]. Such homogenization risks erasing diversity in age, body size, ability, and ethnic features, mirroring historical processes where global media displaced local aesthetics. Addressing these issues requires combining computational audits with ethnographic insights to prevent generative AI from fostering an aesthetic monoculture.

5 Conclusion

CraftGraffiti represents a significant step forward in the intersection of generative AI and urban art, enabling the creation of personalized graffiti that preserves facial identity while embracing stylistic transformation. By introducing a face-consistent self-attention mechanism and leveraging LoRA-based fine-tuning, the system ensures that key facial features such as the eyes, nose, and mouth remain consistent across pose and style changes, addressing the longstanding challenge of identity preservation in diffusion-based image generation. This technical innovation allows for high-quality graffiti renderings that are both aesthetically compelling and faithful to the original subject, blending artistic stylization with personal recognizability. Beyond its technical contributions, *CraftGraffiti* holds broader societal and artistic implications by democratizing access to street art and empowering individuals and communities to participate in shaping their visual environments. By maintaining cultural motifs and individual likenesses, the model enables the creation of meaningful, identity-rich urban expressions that reflect diverse human experiences. As a result, *CraftGraffiti* not only advances the state of the art in generative customization but also offers a transformative tool for preserving and celebrating cultural narratives through the accessible medium of AI-generated graffiti.

Acknowledgements: This piece of research was carried out with the support of the following granted projects: SGR Grant 2021 SGR 01559 from the Catalan Government, GRAIL PID2021-126808OB-I00 and from FEDER/UE, SUKIDI PID2024-157778OB-I00 grants from the Spanish Ministry of Science and Innovation, with the support of Cátedra UAB-Cruïlla grant TSI-100929-2023-2 from the Ministry of Economic Affairs and Digital Transformation of the Spanish Government.

References

[1] Living lab. https://en.wikipedia.org/wiki/Living_lab, 2025. Accessed: 2025-08-09.

[2] Mousa Al-kfairy, Dheya Mustafa, Nir Kshetri, Mazen Insiew, and Omar Alfandi. Ethical challenges and solutions of generative ai: An interdisciplinary perspective. *Informatics*, 11(3):58, 2024. Cited 195 times.

[3] Ayan Banerjee, Josep Lladós, Umapada Pal, and Anjan Dutta. Talediffusion: Multi-character story generation with dialogue rendering. *arXiv preprint arXiv:2509.04123*, 2025.

[4] Lindsay Bates. *Bombing, tagging, writing: An analysis of the significance of graffiti and street art*. PhD thesis, University of Pennsylvania, 2014.

[5] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2017 FMML Workshop on Fair ML*, 2017. arXiv preprint arXiv:1712.03586.

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[8] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Upgpt: Universal diffusion model for person image generation, editing and pose transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4173–4182, 2023.

[9] Luca Compagnucci, Francesca Spigarelli, Jorge Coelho, and Carlos Duarte. Living labs and user engagement for innovation and sustainability. *Journal of Cleaner Production*, 317:128223, 2021.

[10] Saday Chandra Das. Power of graffiti: Exploring its cultural and social significance. *Aayushi International Interdisciplinary Research Journal (AIIRJ), X (IX)*, pages 34–35, 2023.

[11] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023.

[12] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in style. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 2251–2261, 2023.

[13] Emilio Ferrara. Fairness and bias in artificial intelligence: A survey. *Digital*, 6(1):1–41, 2023.

[14] Alan M Forster, Samantha Vettese-Forster, and John Borland. Evaluating the cultural significance of historic graffiti. *Structural Survey*, 30(1):43–64, 2012.

[15] Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. "i don't see myself represented here at all": User experiences of stable diffusion outputs containing representational harms across gender identities and nationalities. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, volume 7, pages 463–475, 2024.

[16] XAI Grok. beta—the age of reasoning agents, 3.

[17] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6986–6996, 2024.

[18] Mark Hamazaspyan and Shant Navasardyan. Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 797–805, 2023.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[20] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[21] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

[23] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023.

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.

[25] Anant Khandelwal. Reposedm: Recurrent pose alignment and gradient guidance for pose guided image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2024.

[26] Suzanne Kieffer. Ecoval: Ecological validity of cues and representative design in user experience evaluations. *AIS Transactions on Human-Computer Interaction*, 9(2):149–172, 2017.

[27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[29] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338, 2019.

[30] Zhen Li, Ping Wang, Qiong Hu, and Ran He. Global and local consistent age generative adversarial network (glca-gan). In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 305–313, 2018.

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

[32] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012.

[33] Cristian Muñoz, Nicola Zannone, Mohamed Mohammed, and Adriano Koshiyama. Uncovering bias in face generation models. *arXiv preprint arXiv:2302.11562*, 2023.

[34] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9192–9201, 2024.

[35] Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

[36] Yuta Okuyama, Yuki Endo, and Yoshihiro Kanamori. Diffbody: Diffusion-based pose and shape editing of human images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6333–6342, 2024.

[37] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. *arXiv preprint arXiv:2502.18461*, 2025.

[38] Rubén Pascual, Adrián Maiza, Mikel Sesma-Sara, Daniel Paternain, and Mikel Galar. Enhancing dreambooth with lora for generating unlimited characters with stable diffusion. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[41] Evangelos Pournaras, Atif Nabi Ghulam, Renato Kunz, and Regula Hänggli. Crowd sensing and living lab outdoor experimentation made easy. *arXiv preprint arXiv:2107.04117*, 2021.

[42] Mohammad Ali Rezaei, Helia Hajikazem, Saeed Khanehgir, and Mahdi Javanmardi. Training-free identity preservation in stylized image generation using diffusion models. *arXiv preprint arXiv:2506.06802*, 2025.

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[45] Chien Sun, William Tzeng, et al. Smiling women pitching down: Auditing gender bias in image generative ai. *arXiv preprint arXiv:2305.10566*, 2023.

[46] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024.

[47] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.

[48] Jiajun Wang, Morteza Ghahremani Boozandani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. *Advances in Neural Information Processing Systems*, 37:65670–65698, 2024.

[49] Quan Wang, Yanli Ren, Xinpeng Zhang, and Guorui Feng. Interactive image style transfer guided by graffiti. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6685–6694, 2023.

[50] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human efforts. *International Journal of Computer Vision*, pages 1–22, 2024.

[51] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7677–7689, 2023.

[52] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023.

[53] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[54] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6383–6392, 2021.

[55] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4442–4451, 2019.

[56] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[57] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023.

[58] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6027–6037, 2023.

[59] Xiang Zhou. Bias in generative ai. *arXiv preprint arXiv:2403.02726*, 2024.

[60] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A Preliminaries

Diffusion Models: Text-to-image diffusion models, such as Stable Diffusion [43, 40, 12], generate images by learning a denoising process that maps latent noise to realistic images ($\mathbf{I}$), conditioned on textual input. Let $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times C}$ denote a clean image and $\mathbf{x}_t$ its noisy version at timestep $t \in \{1, \ldots, T\}$. The forward diffusion process ($q(\cdot)$) adds Gaussian noise ($\mathcal{N}(\cdot)$) in a Markov chain:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

with a predefined variance schedule $\{\beta_t\}$. The model learns the conditional reverse process (p_θ):

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t, \mathbf{c}), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t, \mathbf{c})), \tag{2}$$

where $\mathbf{c}$ is a conditioning vector, often a CLIP-encoded text prompt. Training minimizes the error between predicted and true noise or the original image $\mathbf{x}_0$, depending on parameterization (mean: $\boldsymbol{\mu}_\theta$, and covariance $\boldsymbol{\Sigma}_\theta$). Latent diffusion models (LDMs) [52, 34, 58] perform this process in a compressed latent space $\mathcal{Z}$, using an encoder $\mathcal{E} : \mathbb{R}^{H\times W\times C} \to \mathcal{Z}$ and decoder $\mathcal{D} : \mathcal{Z} \to \mathbb{R}^{H\times W\times C}$, improving computational efficiency while preserving generation quality.

Style Transfer: Style transfer [49, 18, 12] aims to modify the appearance of a generated image $\mathbf{x}$ to match the style of a reference image $\mathbf{s}$, while maintaining the content from a source image $\mathbf{c}$. In feature space $\phi(\cdot)$, the style and content losses are defined as:

$$\mathcal{L}_{\text{style}}(\mathbf{x}, \mathbf{s}) = \sum_l \left\| G^{(l)}(\phi^{(l)}(\mathbf{x})) - G^{(l)}(\phi^{(l)}(\mathbf{s})) \right\|_F^2, \tag{3}$$

$$\mathcal{L}_{\text{content}}(\mathbf{x}, \mathbf{c}) = \left\| \phi^{(l)}(\mathbf{x}) - \phi^{(l)}(\mathbf{c}) \right\|_2^2, \tag{4}$$

where $G^{(l)}$ denotes the Gram matrix [28] of features at layer l, capturing second-order statistics. The total loss combines both objectives:

$$\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_{\text{content}} + \lambda_s \mathcal{L}_{\text{style}}, \tag{5}$$

with tunable weights λ_c and λ_s. Modern approaches integrate style conditioning into generative models via cross-attention modulation or adapter modules. We fine-tune the transformer of the FLUX.1 dev [2] with low-rank adaptation via the $\mathcal{L}_{\text{total}}$ in order to learn the graffiti style and use it as a pre-trained model during inference to pose the style over the input image (see Fig. 2).

Pose Customization with LoRA: Generation of characterization conditions on a specified human pose $\mathbf{p}$, often represented as a 2D keypoint skeleton or heatmap [48, 23, 25]. This signal can be embedded and injected into a diffusion model via attention layers or early-stage concatenation. LoRA [19] offers an efficient means of adapting large diffusion models to new tasks (e.g., pose-guided generation) by injecting low-rank updates into pretrained weight matrices. For a weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA introduces trainable low-rank matrices $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times k}$, modifying the weights as:

$$\tilde{W} = W + \alpha \cdot AB, \tag{6}$$

where α is a scaling factor and $r \ll \min(d, k)$. By freezing base weights and only training A and B, LoRA allows efficient fine-tuning with significantly fewer parameters. LoRA can be applied to attention matrices W_q, W_k, W_v, enabling fast pose-driven customization without full model retraining. In *CraftGraffiti* instead of using keypoints or heatmap, we extend the input prompt via T5 [35] and get the cross-attention map as a guided signal (B) for the pose (see Fig. 2). We maintain the identity of the character through the face-consistent self-attention (A) and update the self and cross attention of the diffusion transformer via fine-tuning with LoRA for multi-scale latent feature generation. This multi-scale latent feature is further denoised via VAE in order to get the final edited image in graffiti style with an accurate musical pose described in the text prompt.

B Ablation Studies

In order to understand the importance of style blending before the pose customization and the face-consistency self-attention, we have performed a brief ablation study depicted in Fig. 6. It has

[2] https://huggingface.co/black-forest-labs/FLUX.1-dev

been observed that with the FLUX.1 dev (12B) baseline, we neither achieve consistency nor achieve the style transfer. On the other hand, we have applied the face consistency self-attention before applying the style transfer. It has been observed that it can preserve the facial attribute during pose customization with low-rank adaptation (2nd row of Fig. 6). Similarly, if we perform style transfer after the pose customization, neither the facial attributes are preserved, even with the presence of face-consistent self-attention, nor does the style reflect graffiti. As depicted in the 3rd row of Fig. 6, the images look like a cartoonish drawing on a wall rather than a graffiti poster. So the correct paradigm is to perform the style blending first and then pose customization with face consistent self-attention (4th row of Fig. 6).

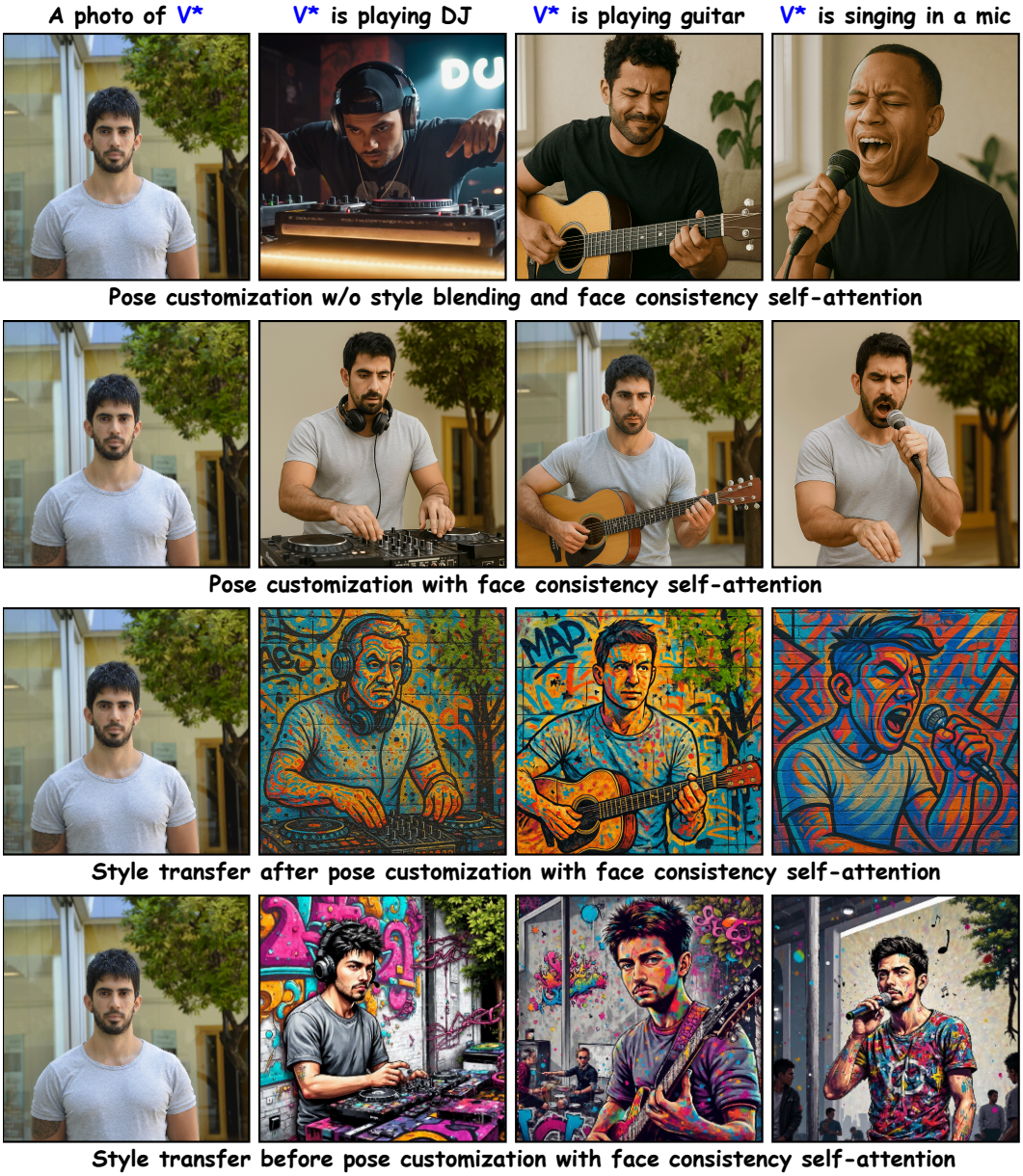

Figure 6: **Ablation Studies:** With face consistency self-attention, we preserve facial attributes during pose customization, and with primary style blending, the graffiti posters appear more realistic.

Similarly, we perform an ablation study on the self-attention (see Fig. 7) to preserve the facial attribute during pose customization to understand why the extension of the attention dimension helps to preserve the facial features. In order to do that, we took the generated images with *CraftGraffiti* and pass them through the self-attention layer of FLUX.1 dev, Subject-driven self-attention layer of

Consistory [46], and our face consistent self-attention layer. It has been observed that the traditional self-attention of FLUX.1 dev focuses on a certain point, whereas Subject-driven Self-Attention of ConsiStory [46] attends the global features rather than the local ones. On the other hand, our proposed face-consistent self-attention, focused on the local features of the facial attributes, helps to maintain facial attributes during face customization.

C Implementation Details

We implement *CraftGraffiti* in PyTorch [39], utilizing a pre-trained .1 dev [3] (12 billion parameter) transformer specifically designed for graffiti style transfer. The guidance scale is set to 7.5. For preserving facial attributes while fine-tuning the transformer with low rank adaptation, the subject guidance factor (λ) is set to 0.95, and the style intensity factor is set to 0.7. We set the rank (σ) to 64 and the regularization factor (α) to 128 while fine-tuning the transformer for pose customization. We run the denoising process with 100 steps by default. We only perform latent composition in the first quarter of the denoising process (the first 25 steps). All experiments ran on a single NVIDIA A6000 48GB GPU.

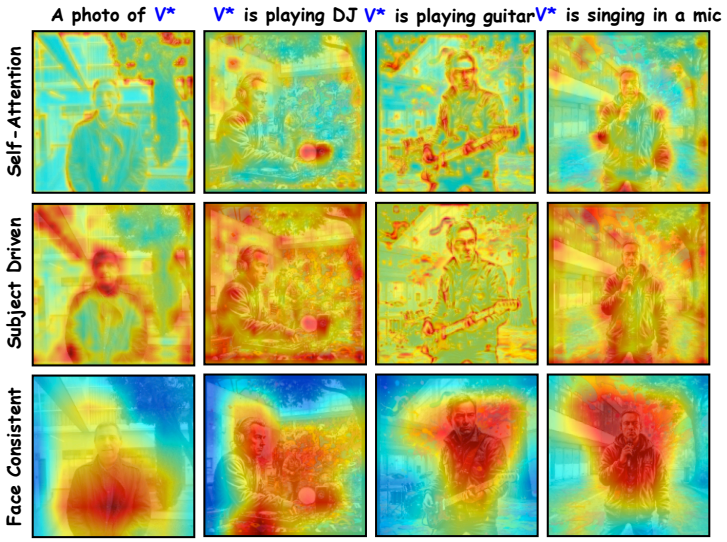

Figure 7: **Ablation of the self-attention:** The face consistent self-attention primarily focuses on the human faces and their corresponding poses, whereas the traditional self-attention and subject-driven self-attention of Consistory [46] diverges towards the global scenario.

Dataset and Evaluation Metric: For graffiti style generation, we have used the images of the 17K-Graffiti dataset [4] to fine-tune our style fusion diffusion with LoRA and use it as a pre-trained model during inference. Also, to further validate *CraftGraffiti*, we have gathered images of people from our laboratory, the Computer Vision Center, Barcelona [5], with their consent, and use them in the paper for validation and pose customization. Furthermore, to perform a fair qualitative analysis, we test *CraftGraffiti* and other state-of-the-art approaches with facial feature consistency (FFC) (implementation in the Appendix), aesthetic score (Aes) [32] calculated using LAION aesthetic classifier, and Human Preference Score (HPS) [53]. Also, we have compared them against the inference time to understand their real-time use case scenario.

D Demonstration at Cruïlla Festival

An installation (see Fig. 8) implementing the proposed system was deployed during the Festival Cruïlla (www.cruillabarcelona.com) in Barcelona from 9-12 July 2025. The festival hosts 80,000 people during 4 days, attending parallel contexts and activities related to culture and innovation. More than 1,100 people visited the booth in which the installation was set, and 586 people were able to create their impression of a graffiti-style poster (see Fig. 9) and give their feedback anonymously. We have summarized the human feedback and listed it as follows:

(1) The general impression of the users was that the demo was fun and engaging, often reacting with surprise and amusement at the results. The majority of participants understood that the demonstration was intended as a playful, exploratory experience rather than a precise or professional tool.

[3] https://huggingface.co/spaces/black-forest-labs/FLUX.1-dev
[4] https://github.com/visual-ds/17K-Graffiti
[5] https://www.cvc.uab.es/

Figure 8: ***CraftGrafitti* artsitic installation:** Conceptual rendering and Actual setup at Cruïlla Festival 2025

(2) Some users asked if their photos were being stored or used to train the algorithm, highlighting the importance of transparency in data management. We informed that none of the user's data has been stored, referring to the informed consent regarding the use of the images only for quality control. The output images of Fig. 9 are are obtained with the permission of the user to showcase the output of the generation for solely research purpose.

(3)The system tended to apply heavy makeup to women's faces, regardless of whether the user was wearing any in the original photo. Generated images of women occasionally included exaggerated features, such as full lips and prominent cheekbones.

(4) A large number of participants commented that the results looked quite similar and their identity was reflected by the system. This was also consistent when asked about their opinion on how the system represented other participants, across users, with only slight variations in hairstyle or clothing, reducing the perceived uniqueness.

(5) The system tended to make people look younger, except for the younger people, in which an explicit evolution on the perceived maturity was identified, particularly in aspects related to body shape, such as muscular development for men or stylized forms for women.

(6) Even though gender consistency is high, for a limited number of cases, some women subjects were interpreted as men. In this case, the user was immediately addressed by the installation operators to validate the reaction and re-state the experimental nature of the installation, addressing the particular misfunctions that it can have and providing valuable feedback for the improvement.

Figure 9: Example outcomes from the demonstration at Cruilla Festival Barcelona 2025

E Some more qualitative examples

We generated some more qualitative examples with *CraftGraffiti* as reported in Fig. 10. It has been observed that, if the user uses some external facial accessories (e.g., classes) *CraftGraffiti* also preserves it in graffiti style (1st and 2nd row of Fig. 10). Similarly, if the user is bald or wearing hijab, it maintains that hair pattern and the external clothing accessories too (3rd and 4th row of Fig. 10).

Figure 10: Some more qualitative examples generated with *CraftGraffiti* maintaining facial attribute with proper style blending.

F Future Directions

Future work is definitly needed, and potential lines or work will have to explore: (1) extending *CraftGraffiti* to handle a broader range of cultural art forms beyond graffiti; (2) integrating real-time bias detection and mitigation pipelines; (3) testing in varied ecological environments, including community art events and educational settings; and (4) expanding cross-cultural studies to assess how identity preservation and stylistic adaptation are perceived in different cultural contexts.

G Ethics Statement

The application of *CraftGraffiti* in image editing offers extensive potential for diverse downstream applications, facilitating the adaptation of images to different contexts. The primary objective of our model is to automate and streamline this process, resulting in significant time and resource savings. It is important to acknowledge that current methods have inherent limitations, as discussed in this paper. However, our model can serve as an intermediary solution, expediting the creation process and offering valuable insights for further advancements. It is crucial to remain mindful of potential risks associated with these models, including the dissemination of misinformation, potential for abuse, and introduction of biases. Broader impacts and ethical considerations should be thoroughly addressed and studied in order to responsibly harness the capabilities of such models. Terms and conditions forms were explicitly offered to all the participants during the live festival, and researchers and support team were always present during the realization of the images, vigilant to potential issues appearing and aware of the user's reactions.

H Code for facial consistency evaluation

Through the following implementation, we first load the FaceNet [44] model pretrained on VGGFace2 [7], which is designed to produce numerical vector embeddings for faces. Given two embedding vectors

$$\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$$

the cosine similarity is defined as:

$$\text{cosine\_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2}$$

where:

- $\mathbf{u} \cdot \mathbf{v}$ is the dot product of the two vectors.

- $\|\mathbf{u}\|_2$ and $\|\mathbf{v}\|_2$ are their Euclidean norms.

A value close to 1 indicates that the two embeddings are highly similar in direction, meaning the corresponding faces are likely to be of the same identity. A value near 0 indicates no similarity, while negative values imply opposite feature orientations.

Listing 1: Computing cosine similarity between two face embeddings

```python
from PIL import Image
import numpy as np
import torch
from facenet_pytorch import InceptionResnetV1
import torchvision.transforms as T
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt

# Load FaceNet model
model = InceptionResnetV1(pretrained='vggface2').eval()

# Load and preprocess images
transform = T.Compose([
    T.Resize((160, 160)),
    T.ToTensor(),
    T.Normalize([0.5], [0.5])
])

img1 = transform(Image.open("face1.png")).unsqueeze(0)
img2 = transform(Image.open("face2.png")).unsqueeze(0)

# Get embeddings
emb1 = model(img1).detach().numpy()
emb2 = model(img2).detach().numpy()

# Cosine similarity
similarity = cosine_similarity(emb1, emb2)[0][0]
print(f"Cosine similarity: {similarity:.4f}")
```

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes] .

   Justification: We proposed a diffusion-based text-to-image customization model by solving the curse of dimensionality of the attention mechanism. The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper.

   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]

   Justification: We have created a limitations section stating the limitations of the proposed work.

   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: [Yes]

Justification: All the theorems, formulas, and proofs in the papers are numbered and cross-referenced

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: We provide a GitHub link to have access to the code.

   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: We provide a GitHub link to have access to the code.

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: We have written the implementation details for this.

   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: [Yes]

   Justification: Please refer to the experimental analysis.

   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: We indicate the type of compute workers, CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

   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: We preserve all the anonymity.

   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 reflects the impact of art on Humanity.

    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 the 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:

    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: As per our best knowledge, all the data, code, and models are original.

    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: [NA]

    Justification:

    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: [NA]

    Justification:

    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: [NA]

    Justification:

    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: [Yes]

Justification: We use LLM to generate the possible prompt that correctly reflects the graffiti style fed into the CLIP text encoder.

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.