

GenEAvA: Generating Cartoon Avatars with Fine-Grained Facial Expressions from Realistic Diffusion-based Faces

Hao Yu¹, Rupayan Mallick², Margrit Betke¹, Sarah Adel Bargal²

¹ Department of Computer Science, Boston University, Boston, USA

² Department of Computer Science, Georgetown University, Washington, D.C., USA

Abstract— Cartoon avatars have been widely used in various applications, including social media, online tutoring, and gaming. However, existing cartoon avatar datasets and generation methods struggle to present highly expressive avatars with fine-grained facial expressions and are often inspired from real-world identities, raising privacy concerns. To address these challenges, we propose a novel framework, GenEAvA, for generating high-quality cartoon avatars with fine-grained facial expressions. Our approach fine-tunes a state-of-the-art text-to-image diffusion model to synthesize highly detailed and expressive facial expressions. We then incorporate a stylization model that transforms these realistic faces into cartoon avatars while preserving both identity and expression. Leveraging this framework, we introduce the first expressive cartoon avatar dataset, GenEAvA 1.0, specifically designed to capture 135 fine-grained facial expressions, featuring 13,230 expressive cartoon avatars with a balanced distribution across genders, racial groups, and age ranges. We demonstrate that our fine-tuned model generates more expressive faces than the state-of-the-art text-to-image diffusion model SDXL. We also verify that the cartoon avatars generated by our framework do not include memorized identities from fine-tuning data. The proposed framework and dataset provide a diverse and expressive benchmark for future research in cartoon avatar generation.

I. INTRODUCTION

Cartoon avatars have become increasingly important in various digital domains, serving as personalized digital representations in applications such as social media [41], chatbots [22], online tutoring [18], [57], video conferencing [45], virtual reality [5], and video games [40]. As digital communication evolves, cartoon avatars offer a compelling alternative to realistic human representations, providing users with enhanced personalization and privacy, and enriching user engagement and interaction across various platforms.

Despite the growing popularity of cartoon avatars across various applications, current avatar generation methods and cartoon face datasets have several limitations. Many existing approaches struggle to create highly expressive cartoon avatars and fail to effectively convey nuanced emotions [59]. This is partly due to the lack of cartoon face data with diverse facial expressions, as most available datasets predominantly feature neutral or basic facial expressions [21]. Furthermore, generative models sometimes memorize identities or generate avatars that resemble real individuals from the training data rather than producing genuinely novel identities [62], which raises significant privacy concerns. An ideal avatar generation system should create diverse and unique representations without overfitting to specific individuals. Additionally, cartoon face datasets often exhibit bias in terms of age and race,

with many skewed toward young, lighter-skinned characters, e.g., Manga109 [19].

In light of these challenges, we propose a novel framework, GenEAvA, for Generating Expressive cartoon Avatars. In this work, we specifically define cartoon avatars as cartoon-style digital representations that represent identities in the images generated by the diffusion model. We first propose a facial expression generation model that can generate fine-grained realistic facial expressions across 135 emotion categories based on fine-tuning a text-to-image (T2I) diffusion model. High-quality realistic facial expressions are then generated using this model with carefully designed text prompts. The prompts are curated using the state-of-the-art Large Language Model (LLM) GPT-4o [27], ensuring a wide range of age groups, equal representation of males and females, and a balanced racial distribution across seven racial groups. Finally, we convert these realistic facial expressions into cartoon avatars through a stylization method while maintaining the identity and facial expression of the original faces. Additionally, we present a comprehensive evaluation pipeline for cartoon avatar generation, focusing on facial expression fidelity and representation, identity memorization, and the preservation of identity and expression during stylization.

We present GenEAvA 1.0, the first cartoon avatar dataset that is specifically designed to include fine-grained facial expressions with unique identities, diverse age groups, and a balanced racial distribution. GenEAvA 1.0 consists of 13,230 cartoon avatars of 135 facial expressions. We conducted extensive experiments to evaluate the quality of the cartoon avatars in the dataset and show that the generated images in GenEAvA 1.0 present fine-grained facial expressions, surpassing the state-of-the-art T2I diffusion model SDXL [48] across various visual quality metrics. We also demonstrate that the dataset includes novel identities without instances of memorization from the fine-tuning dataset through quantitative analysis and a user study. Finally, we validate that fine-grained facial expressions and novel identities are maintained through the stylization module using both quantitative analysis and a user study.

We summarize our contributions as follows:

- We propose a novel framework, GenEAvA, for the generation of expressive cartoon avatars from realistic faces generated by T2I diffusion models.
- We fine-tune the state-of-the-art T2I SDXL diffusion model to generate particularly fine-grained facial expressions.

- We propose a diverse cartoon avatar dataset, GenEAvA 1.0, with fine-grained facial expressions, unique identities, and balanced age, gender, and racial distribution.

II. RELATED WORK

Our work lies at the intersection of facial expression generation, memorization in generative models, and synthetic face data. We review the related literature for each domain.

A. Facial Expression Generation

Facial expression generation refers to the process of synthesizing or modifying facial expressions in images or videos. Earlier approaches for facial expression generation are based primarily on Generative Adversarial Networks (GANs) [20]. For example, StarGAN [13] enables facial image editing with basic expressions through multi-domain image-to-image translation. To achieve more fine-grained control over facial expressions, several methods have incorporated Action Units (AUs) for more precise expression manipulation [49], [63], [68]. GANmut [14] further learns an expressive and interpretable conditional space of emotions to generate compound emotions. EmoStyle [2] uses StyleGAN2 [29] with a valence-arousal space for intuitive, continuous expression control.

More recently, diffusion models [24] have achieved significant success in generating high-quality images and have been applied to facial expression generation. Stable Diffusion [52] has shown an exceptional ability to generate high-quality images from text prompts with basic facial expressions. Pikoulis *et al.* [47] fine-tuned Stable Diffusion using CLIP latent guidance to generate seven basic emotions. The most relevant work to ours is by Liu *et al.* [34], which queries a dataset of 135 expressions and transfers the specific facial expression using a conditional diffusion model. In contrast, our approach does not rely on selecting a reference image for expression transfer. Instead, we fine-tune the diffusion model to learn the distribution of 135 facial expressions and sample directly from it, which enables the generation of novel images with intended facial expressions.

B. Identity Memorization in Generative Models

Memorization in a generative model refers to the generation or reproduction of the training data by a trained model at the time of inference. Initial empirical studies involving GANs [64] have challenged the novelty of generation using a kind of membership inference attack. Later, generative models such as diffusion models have also been shown to replicate the underlying training priors at the time of generation. Somepalli *et al.* [60] identified direct replication of training data by stable diffusion. Furthermore, Carlini *et al.* [8] demonstrated the lack of privacy preservation in the diffusion models, showing the leak in the training data at the inference time. Carlini further showed that replication can be induced by prompting the image captions from the LAION dataset, which is used for training diffusion models. Several works studied the mitigation of memorized models [11], [61] to reduce the regeneration of training data. Other

methods include differentially private generative models for privacy-preserving image generation methods [4], [10], [17]. In this work, we focus on identity memorization, which is a specific type of memorization in generative models where the generated face images replicate real identities from the training data.

C. Synthetic Face Data

Recent advances in generative models (*e.g.*, Generative Adversarial Networks [20], [55], Diffusion Models [24], [52]) have enabled the creation and use of synthetic face data for face-related computer-vision tasks such as face recognition, landmark localization, and face parsing [3], [6], [30], [36], [50], [65]. Synthetic face datasets like DigiFace-1M [3], the Face Synthetics Dataset [65], DCFace Synthetic Dataset [30], and the GANDiffFace Dataset [36] provide millions of synthetic faces with diverse attributes, which serve as alternatives to real-world datasets, mitigating privacy and bias issues. However, these synthetic face datasets are mostly designed for face detection and recognition, and include only limited facial expressions.

Studies have also explored the use of synthetic data for facial expression recognition [1], [39], [23]. For example, Abbasnejad *et al.* [1] used a 3D face model to generate six basic facial expressions. SynFER [23] creates synthetic facial expression data with generative models based on facial action units and text guidance. Although these studies limit themselves to six basic facial expressions, we present a pipeline and dataset of 135 fine-grained facial expressions.

Several datasets have been specifically developed for cartoon faces. IIIT-CFW [38] contains 8,928 cartoon faces of 100 public figures. WebCaricature [26] includes 6,042 caricatures and 5,974 photographs annotated with 17 facial landmarks. Manga109 [19] is a collection of 109 Japanese comic books designed for face detection tasks, and Danbooru [7] comprises over 970k anime images from 70k identities. Cartoon Set [21] provides sets of 10k and 100k 2D cartoon avatar images. The iCartoonFace Dataset [70] includes 389,678 images of 5,013 cartoon characters annotated for face detection and recognition. To the best of our knowledge, the cartoon face dataset that we introduce is the first dataset that specifically focuses on presenting cartoon face images with fine-grained facial expressions across 135 categories.

III. METHOD

We propose GenEAvA, a novel pipeline for generating high-quality cartoon faces with fine-grained facial expressions. As shown in Figure 1, the pipeline includes two phases: fine-tuning and inference. We first build a facial expression generation model that can generate photorealistic faces with fine-grained facial expressions based on a state-of-the-art text-to-image (T2I) diffusion model. We finetune a pretrained T2I model on 135 classes of facial expressions using both a diffusion model loss and an expression loss. At inference, we generate high-quality face images with diverse facial expressions by prompting the T2I model and then

utilize a stylization model to convert these photorealistic faces into cartoon avatars.

A. Preliminary on Text-to-Image Diffusion Models

Text-to-Image (T2I) diffusion models represent an emerging class of generative models that have recently achieved impressive results on various generative modeling benchmarks [54], [66]. They consist of a forward process and a reverse process [24]. In the forward process, a data sample \mathbf{x}_0 is incrementally converted into pure Gaussian noise over a series of diffusion steps, where Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ is gradually added at each step t , resulting in a sequence of intermediate noisy images \mathbf{x}_t . The reverse diffusion process begins with a standard Gaussian distribution and iteratively removes noise to generate a sample that resembles the training distribution. For training efficiency, these processes are always operated in a latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$ obtained using an image encoder \mathcal{E} . Given a timestep t , an intermediate noisy latent feature \mathbf{z}_t , and a text feature $\tau(p)$, generated by a text encoder τ and a text prompt p , a T2I diffusion model trains a conditional denoising U-Net [53] to predict the added noise ϵ using the squared error loss

$$\mathcal{L}_{\text{dm}} = \|\epsilon - \epsilon_{\theta}(\mathbf{z}_t, t, \tau(p))\|_2^2. \quad (1)$$

The closed form of \mathbf{z}_t can be derived as:

$$\begin{aligned} \mathbf{z}_t &= \sqrt{\bar{\alpha}_t} \mathbf{z}_0 + (1 - \bar{\alpha}_t) \epsilon, \\ \text{with } \bar{\alpha}_t &= \prod_{s=1}^t \alpha_s \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, 1), \end{aligned} \quad (2)$$

where α_s is a sequence of predefined coefficients controlling the variance of noise added at each step. Trained on large-scale image-text datasets such as LAION-5b [56], T2I diffusion models are capable of generating high-quality images from text descriptions.

B. Expression-guided Finetuning

To enable the T2I diffusion model to generate accurate, diverse, and fine-grained facial expressions, we finetune a state-of-the-art pre-trained T2I diffusion model, SDXL [48], on in-the-wild facial expression images.

a) *Finetuning on Diverse Expression Images*: We adopt Emo135 [12], a face dataset containing 135 fine-grained expression categories. For the best generation quality, we perform a series of pre-processing steps. First, we detect and crop faces in the images using the face detector RetinaFace [16]. Since some images in the dataset contain watermarks, we use a watermark detection algorithm [46] to filter out such images. Additionally, to address the imbalance of facial expressions in the original dataset, we construct a balanced dataset of 135 facial expressions by randomly sampling from the original dataset. This process results in a curated dataset containing 1,080 images, which is then used to finetune the SDXL model.

b) *Training objectives*: While standard diffusion models are trained using a squared loss as in Eq. 1, we incorporate an expression loss \mathcal{L}_{exp} to guide the model toward generating more accurate and nuanced facial expressions. We use an expression encoder \mathcal{E}_{exp} to extract the expression representation of an image. Specifically, we employ a state-of-the-art facial expression recognition model, POSTER [69], as our expression encoder. To guide the generation process, we compute the mean squared error (MSE) between the expression representations of the generated image $\hat{\mathbf{x}}_0$ and the real image \mathbf{x}_0 . The expression loss is defined as:

$$\mathcal{L}_{\text{exp}} = \text{MSE}(\mathcal{E}_{\text{exp}}(\mathbf{x}_0), \mathcal{E}_{\text{exp}}(\hat{\mathbf{x}}_0)). \quad (3)$$

As derived previously [24], we can approximate $\hat{\mathbf{x}}_0$ at any timestep t by a one-step reverse formula from Eq. 2, which is defined as:

$$\begin{aligned} \hat{\mathbf{z}}_0 &= \frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}}{\sqrt{\bar{\alpha}_t}}, \\ \hat{\mathbf{x}}_0 &= \mathcal{D}(\hat{\mathbf{z}}_0), \end{aligned} \quad (4)$$

where \mathcal{D} is an image decoder. The overall training objective is a combination of the original diffusion model loss and the expression loss with an α scaling factor, formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{dm}} + \alpha \mathcal{L}_{\text{exp}}. \quad (5)$$

C. Facial Expression Generation

Given the fine-tuned SDXL model, we generated high-quality facial expression images through carefully designed prompts. These prompts were specifically tailored to ensure a balanced representation of gender, age, and race in the generated images. We leveraged the state-of-the-art large language model GPT-4o [27] to generate prompts that include equal representation of males and females, cover a wide range of ages from teenagers to elderly individuals, and provide a balanced representation across seven racial groups (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino [28]). A sample prompt is: “A photorealistic face of a middle-aged Indian woman with shoulders visible, displaying a facial expression of delight, plain white background.”

To ensure consistent image quality, we filtered out images where the face appears too close to the camera, resulting in incomplete facial features, as well as images containing multiple generated faces.

D. Cartoon Style Transfer

Lastly, a face stylization method is applied to convert the generated images into a cartoon style. The framework allows for the integration of alternative stylization methods, and the evaluation pipeline presented in Section V can still be applied for assessing the entire framework. In this study, we use DCTNet [37], a state-of-the-art image translation architecture for few-shot portrait stylization. We selected it for its advanced ability to synthesize high-fidelity content and its strong generality. It was trained to synthesize artistic portraits in various styles. We employed its pre-trained model for 3D cartoon style to generate cartoon avatars with fine-grained facial expressions. We then validated that fine-grained facial

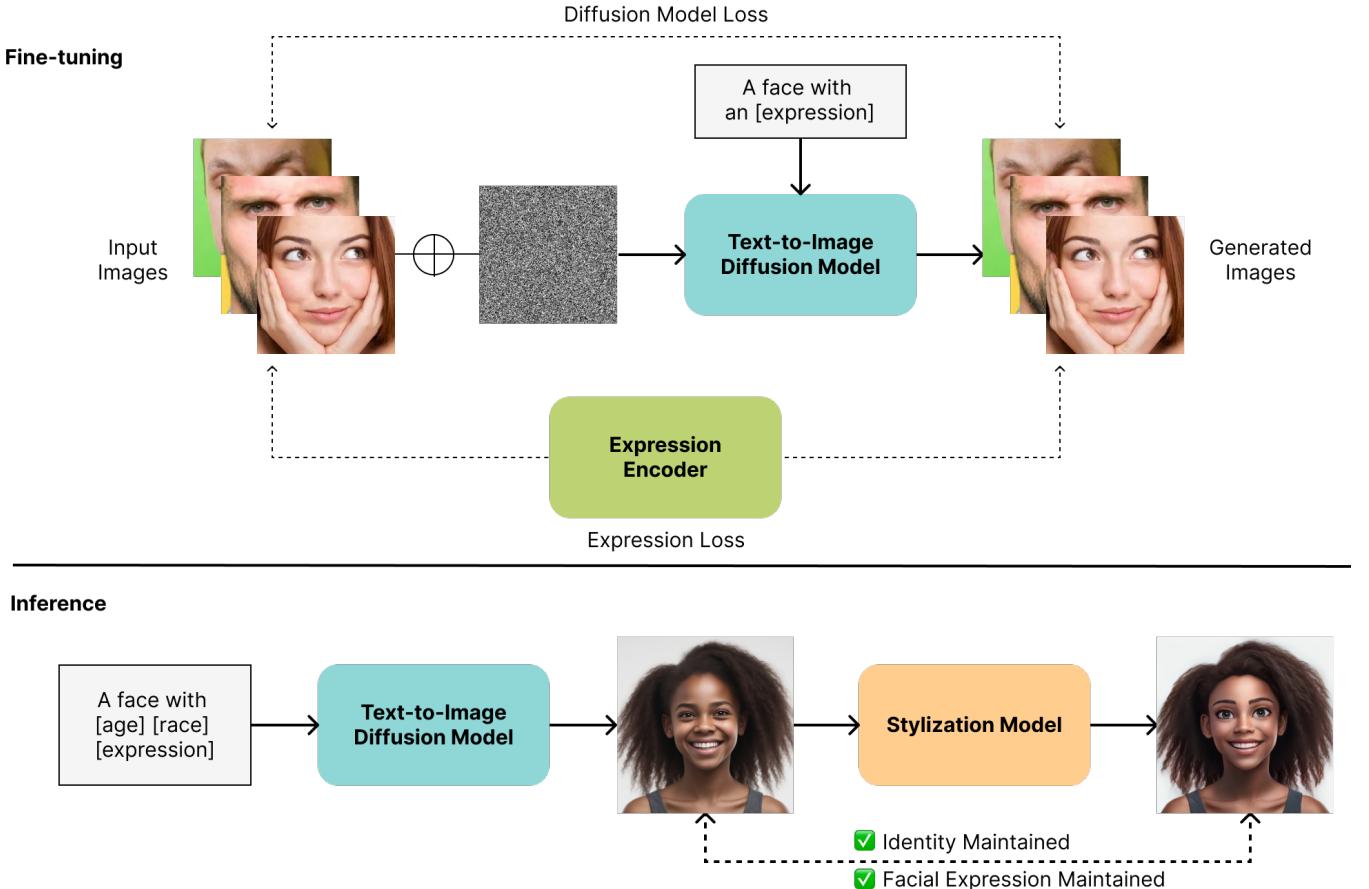


Fig. 1. The proposed Pipeline, GenEAvA, for generating expressive cartoon avatars. During the fine-tuning phase, we train a text-to-image diffusion model using facial expression images. The model is optimized with a combination of diffusion model loss (DM loss) and expression loss computed by an expression encoder. In the inference phase, we generate facial expression images by prompting the model, followed by applying a stylization model to transform them into cartoon avatars.

expressions are maintained through the stylization module using a user study presented in Section V-D.1.

IV. DATASET: GENEAVA 1.0

Building on the pipeline described above, we present GenEAvA 1.0, a novel dataset of **Generative Expressive Avatars**. GenEAvA 1.0 comprises 13,230 cartoon avatars of 135 facial expressions. Each facial expression is represented by 98 images, ensuring a balanced distribution across genders (male and female), seven racial groups, and various age ranges.

Examples of generated photo-realistic faces and their corresponding cartoon avatars are shown in Figure 2. As illustrated in the figure with a sample of 21 different expressions, the dataset includes cartoon avatars with fine-grained facial expressions (135 classes) and represents diverse races, ages, and genders. The images also feature clean backgrounds, further enhancing their utility.

V. EXPERIMENTS

A. Experimental Setup

We conducted experiments for both facial expression generation and stylization to evaluate different aspects of the proposed framework. First, we assessed the performance of

TABLE I
FACIAL EXPRESSION GENERATION RESULTS. OUR MODEL OUTPERFORMS THE SDXL MODEL ACROSS ALL METRICS, INCLUDING CLIP, DINO, LPIPS, AND EXPRESSION ERROR (EXP.) SCORES.

Model	CLIP↑	DINO↑	LPIPS↓	Exp.↓
SDXL [48]	0.780	0.738	0.658	13.1
Ours	0.799	0.742	0.648	12.6

the facial expression generation model in producing images with accurate and fine-grained facial expressions. Second, we investigated potential memorization issues within the model to ensure it generates images with unseen identities and avoids replicating those present in the fine-tuning data. Lastly, we evaluated the stylization model to determine its ability to preserve the original content of the images, focusing specifically on maintaining identity and facial expressions during the stylization process.

B. Facial Expression Generation

We fine-tuned the SDXL model on the Emo135 dataset using LoRA [25] with a rank of 4. We used the official



Fig. 2. Examples of realistic and stylized images across a variety of facial expressions in GenEAvA 1.0. The images illustrate diverse age groups and a balanced representation of race and gender. The stylization effectively preserves the identity and expressions of the realistic images.

SDXL checkpoint from Hugging Face.¹ The learning rate is set to 1e-6. We trained the model for eight epochs with a batch size of 1. More epochs lead to overfitting and worse image quality. The Adam optimizer [31] was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1e-2. The expression loss weight α is set to 1.0. All experiments were conducted on four NVIDIA RTX A6000 GPUs.

To evaluate the model’s ability to generate fine-grained facial expressions, we conducted experiments comparing our model to SDXL, the state-of-the-art T2I diffusion model. A total of 13,230 images were generated using SDXL with the same prompts used to create our proposed dataset GenEAvA. We then computed multiple metrics to assess the fidelity and representation of facial expressions in both the

SDXL-generated images and our dataset. For evaluation, we randomly sampled an evaluation subset from Emo135, consisting of 50 images for each facial expression.

The following are the four metrics used for the discussed assessment: CLIP [51], DINO [9], LPIPS [67], and expression error (Exp.). The CLIP metric measures the average pairwise cosine similarity between the CLIP [51] embeddings of the generated images and the Emo135 evaluation images, capturing semantic consistency between the generated and real facial expressions. The DINO metric is the average pairwise cosine similarity between the DINOV2 [44] embeddings of the generated and evaluated images. LPIPS quantifies low-level perceptual differences, focusing on fine-grained texture and feature fidelity in images. It is calculated as the average pairwise similarity between the AlexNet [32] activations of the generated and evaluation images. Finally,

¹<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

the expression error is the average pairwise Euclidean distance between the expression embeddings [69] of the generated and the evaluation images. We cropped the face in the image using RetinaFace [16] and extracted the expression embedding using the facial expression model POSTER [69].

The results of GenEAvA and the baseline SDXL, measured in terms of the four metrics, are presented in Table I. Our model outperforms SDXL across all the evaluation metrics, indicating the GenEAvA's superior ability to generate fine-grained facial expressions.

C. Identity Memorization in Facial Expression Generation

To evaluate whether the trained facial expression generation model memorizes the identities in the fine-tuning data, we compared the identity embeddings of generated images in GenEAvA 1.0 with those of the images in the fine-tuning dataset. First, we used RetinaFace [16] to extract faces from images and a face recognition model, ArcFace [15], to compute identity embeddings. The cosine similarity between the identity embeddings of the generated images and the fine-tuning images is then calculated. For each generated image, we identified the training image with the highest cosine similarity (the most similar face) and determined whether the two images belong to the same identity on the basis of (1) an empirical verification threshold and (2) a user study. Using an algorithm that maximizes the information gain [58], the threshold is set to 0.68. If the cosine similarity was to exceed this threshold, the two faces would be considered to have the same identity. We also conducted a user study to check whether the generated images replicate identities from previously seen data.

1) Quantitative Analysis: We first compared the generated images with those in our fine-tuning dataset Emo135 [12]. None of the generated faces exceeded the verification similarity threshold [58] for any of the faces in the Emo135 dataset. Additionally, the average pairwise cosine similarity between the generated faces and their most similar counterparts in the fine-tuning set is 0.39, which is significantly below the verification threshold. This suggests that the model did not memorize the identities from the fine-tuning dataset.

Ideally, we would also evaluate this metric on the entire SDXL training dataset. However, since the SDXL training data is not fully publicly accessible, we used CelebA [35], a commonly used face dataset known to be part of SDXL's training data. There are 10,177 distinct identities in the dataset, and we randomly selected one image for each identity for comparison. None of the generated faces exceeds the verification similarity threshold when matched against any of the CelebA faces. Additionally, the average cosine similarity between the embeddings of the generated faces and their closest counterparts in CelebA is 0.47.

2) User Study: To further validate that identities have not been memorized in the generated images, we conducted a user study. We randomly sampled 50 images from GenEAvA 1.0 and found their most similar faces in the fine-tuning dataset Emo135 based on identity similarity. Similarly, we also sampled 50 images for comparison with the CelebA

dataset. Then, five participants were asked to determine whether the paired faces belong to the same person. Out of 100 pairs, two participants identified one pair of faces as the same identity, two participants identified two pairs, and one participant identified five pairs. We note that all these faces are from the CelebA dataset, and the pairs share similarities, but are not exact replications. This suggests that our model did not memorize the identities in the fine-tuning dataset, Emo135. Also, a few similar faces to CelebA could be attributed to SDXL's pretraining on large-scale public datasets, where celebrities are overrepresented. We will remove these faces that are identified as similar to CelebA in the final version of our dataset. We will also apply the latest unlearning algorithm (*e.g.*, [33]) to make the model forget such content.

Overall, these results demonstrate that our model does not memorize identities from the fine-tuning dataset, and that the generated dataset does not include any identities present in these training data. The proposed GenEAvA 1.0 establishes a benchmark with new identities that feature a balanced representation of race, gender, and age, promoting both privacy and fairness.

D. Identity and Facial Expression Preservation in Face Stylization

Finally, we evaluate the performance of face stylization. An ideal stylization method should transform the image into the desired style while preserving the original content that is not related to style. For avatar generation, we specifically focus on maintaining the identity and facial expression of the face. This evaluation was carried out with a user study.

1) User Study: We conducted a user study comparing the realistic fine-grained facial expression image to its corresponding avatar generation in terms of (i) identity preservation, and (ii) facial expression preservation. We used the Amazon Mechanical Turk (AMT) crowdsourcing marketplace² to recruit seven crowd workers for each task. For the user study, we randomly sampled 945 generated image-avatar pairs (seven from each of 135 emotion classes). We recruited AMT workers who had previously completed at least 500 tasks ('HITs'), and maintained an approval rating of at least 95%. We compensated for the work of all crowd workers who participated in our tasks. For each HIT, a total of 9 evaluators were assigned. Each subtask presents the worker with one realistic fine-grained expression image and its corresponding avatar image (Figure 3). The worker is asked to determine whether the identity and fine-grained facial expression are maintained from the realistic generation to the resulting avatar. We posted all HITs simultaneously and allocated a maximum of 9 minutes to complete each HIT. Each evaluator was presented with 15 image pairs, one of which is a test question presenting two images with obviously different identities. This is used to evaluate the validity of the HIT response. Invalid HITs, *i.e.*, those with an evaluator incorrectly answering the test questions, were discarded.

²<https://www.mturk.com/>

Instructions

1. You will see two AI-generated images.
2. Your task is to answer questions about the images.
3. On the left, you will see a realistic looking face. On the right, you will see a cartoon version of the same face.
4. Evaluate the two images based on the two questions below the image pair.



Realistic



Cartoon Avatar

a) Does the avatar represent the identity of the realistic face on the left?

Yes No

b) Does the avatar preserve the facial expression of the person in the realistic image?

Yes No

Fig. 3. Interface for the Amazon Mechanical Turk (AMT) user study. The first question addresses the preservation of identity through the stylization module, and the second question addresses the preservation of the facial expression through the stylization module. Each evaluator was presented with 15 such examples, one of which is a test question presenting two images with obviously different identities. This is used to evaluate the validity of the HIT. Nine Turkers were recruited to complete each HIT. Invalid HITs were discarded.

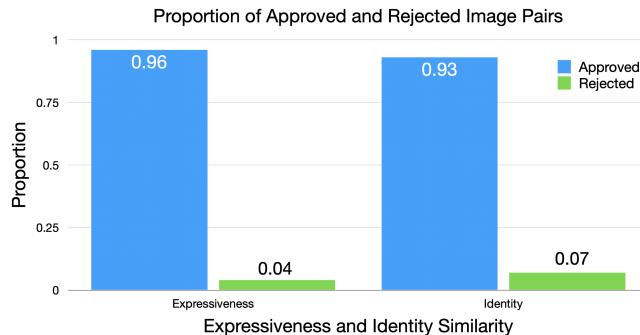


Fig. 4. User study results evaluating the stylization based on identity preservation and expression preservation. We achieved 96% approval rating in preserving facial expression and 93% approval rating in preserving identity, indicating the effectiveness of the stylization method. The approval rating indicates the percentage (%) of pairs that preserve the facial expression and identity among them.

The results show that our dataset achieved 96% approval rating in facial expression preservation and 93% approval rating in identity preservation (Figure 4). This suggests that the stylization method effectively maintains the facial expression and identity of the images. We will remove images that fail to preserve the identity and facial expression from the final version of the dataset.

E. Qualitative Results

We present qualitative examples that compare our cartoon avatars with those generated by ChatGPT [42] using GPT-4o [27] and DALLE-3 [43] in Figure 5. For ChatGPT-generated images, the sample prompt we used is: “*Generate a detailed cartoon avatar of a middle-aged White male showing a facial expression of compassion.*”

As shown in Figure 5, our proposed method GenEAvA captures subtle facial expressions more accurately than ChatGPT. For example, ‘compassion’ is defined as sympathetic pity and concern for the sufferings or misfortunes of others. However, the face generated by ChatGPT depicts a smile while GenEAvA produces a slight frown, which better aligns with the intended expression. For some classes like ‘desire’ and ‘sympathy,’ ChatGPT generates neutral faces, while our proposed method correctly captures the subtle expressions. For ‘jealousy,’ the ChatGPT-generated face is overly exaggerated compared to the proposed GenEAvA. Overall, we show that it is challenging even for commercial multimodel LLMs like ChatGPT to accurately generate certain subtle facial expressions. Our proposed method is able to successfully create faces with fine-grained facial expressions.

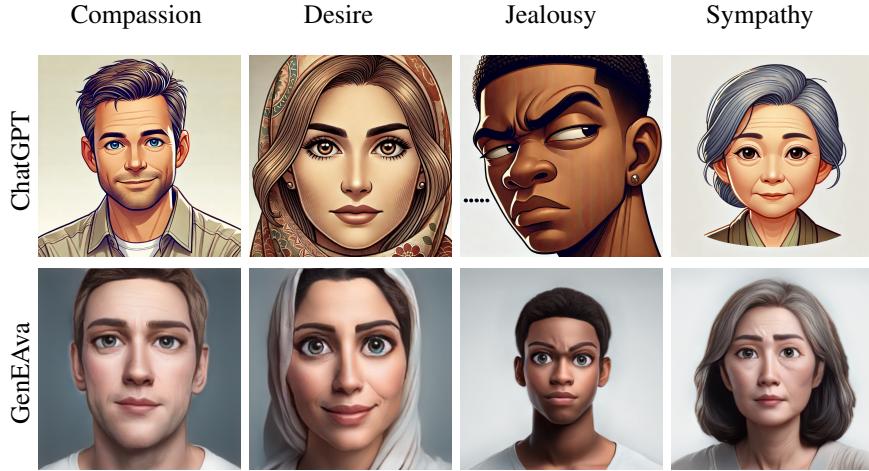


Fig. 5. Qualitative examples of images generated by ChatGPT [42] and our proposed GenEAvA. GenEAvA shows a superior ability to capture subtle expressions compared to ChatGPT, which either produces generic neutral or exaggerated expressions.

VI. CONCLUSIONS

In this work, we presented a novel framework, GenEAvA, which can generate and validate high-quality cartoon avatars with fine-grained facial expressions while ensuring unique identities and diversity in age, gender, and racial representation. We fine-tuned a text-to-image diffusion model with expression-guided objectives to generate highly detailed and expressive facial expressions. We designed comprehensive prompts for the T2I diffusion model to produce images with equal representation of males and females, a wide range of ages from teenagers to elderly individuals, and a balanced representation across seven racial groups (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino). Finally, we applied a stylization model that converts these realistic faces into cartoon avatars while preserving both identity and expression. In addition, we presented an evaluation pipeline for cartoon avatar generation, assessing the fidelity and representation of facial expression, identity memorization, and the preservation of identity and expression in stylization.

Leveraging this framework, we introduced GenEAvA 1.0, the first expressive cartoon avatar dataset specifically designed to capture 135 fine-grained facial expressions, featuring 13,230 cartoon avatars with a balanced distribution across demographic attributes. Extensive experiments demonstrate that our model outperforms state-of-the-art diffusion-based text-to-image models, such as SDXL, in generating subtle and accurate facial expressions. We verified that the generated identities are novel and not memorized from the training set through both quantitative analysis and a user study. We also showed that identity and fine-grained facial expressions are preserved by the stylization method.

Our findings highlight the potential of diffusion-based models in advancing expressive cartoon avatar generation. Future work could explore further improving expression control, improving identity consistency across different expressions, and adapting the framework for real-time applications.

ETHICAL IMPACT STATEMENT

Our work aims at generating cartoon avatars with fine-grained facial expressions. As with all generative models, our model has inherent risks and potential negative impacts, including potential misuse for generating misleading or harmful content, privacy concerns regarding the unauthorized use of individuals' likenesses, and the risk of reinforcing or introducing biases through the generated images. Furthermore, while facial expression generation and manipulation could enhance digital interactions, gaming, and virtual communication, it also poses ethical concerns, such as impersonation and deceptive emotional representations. We acknowledge these risks and emphasize the need for responsible use, transparency, and ethical deployment of the proposed model and AI-generated content. To mitigate these risks, we ensure that there is no identity memorization in our model and dataset through extensive experiments. We also firmly oppose the use of our model for generating fake or misleading content. Future users of our dataset will be required to agree to our terms of use and ethical guidelines.

REFERENCES

- [1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1609–1618, 2017.
- [2] B. Azari and A. Lim. EmoStyle: ne-shot facial expression editing using continuous emotion parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6385–6394, 2024.
- [3] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen. DigiFace-1M: 1 million digital face images for face recognition. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.
- [4] A. Bie, G. Kamath, and G. Zhang. Private GANs, revisited. *Transactions on Machine Learning Research*, 2023. Survey Certification.
- [5] P. Bimberg, M. Feldmann, B. Weyers, and D. Zielasko. The influence of environmental context on the creation of cartoon-like avatars in virtual reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 298–308. IEEE, 2024.

- [6] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. SFace: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–11. IEEE, 2022.
- [7] G. Branwen and A. Gokaslan. Danbooru2019: A large-scale crowdsourced and tagged anime illustration dataset. <https://gwern.net/danbooru2021#danbooru2019>, 2019.
- [8] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium, SEC ’23, USA, 2023*. USENIX Association.
- [9] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [10] C. Chen, D. Liu, S. Ma, S. Nepal, and C. Xu. Private Image Generation with Dual-Purpose Auxiliary Classifier . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20361–20370, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.
- [11] C. Chen, D. Liu, and C. Xu. Towards memorization-free diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8425–8434, 2024.
- [12] K. Chen, X. Yang, C. Fan, W. Zhang, and Y. Ding. Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1906–1916, 2022.
- [13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.
- [14] S. d’Apolito, D. P. Paudel, Z. Huang, A. Romero, and L. Van Gool. GANmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021.
- [15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [16] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, May 2019.
- [17] T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.
- [18] M. C. Fink, S. A. Robinson, and B. Ertl. AI-based avatars are changing the way we learn and teach: benefits and challenges. In *Frontiers in Education*, volume 9, page 1416307. Frontiers Media SA, 2024.
- [19] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa. Manga109 dataset and creation of metadata. In *MANPU ’16: Proceedings of the 1st International Workshop on coMics Analysis, Processing and Understanding*, pages 1–5, 2016.
- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [21] Google. CartoonSet: A dataset of cartoon faces. <https://google.github.io/cartoonset/>, 2021. Accessed: 2025-01-18.
- [22] A. M. Hadjiev and K. Araki. Evaluation of various avatar designs for conversational chatbot systems. *SIG-LSE JSAI*, 2021.
- [23] X. He, C. Luo, X. Xian, B. Li, S. Song, M. H. Khan, W. Xie, L. Shen, and Z. Ge. SynFER: Towards boosting facial expression recognition with synthetic data. *arXiv preprint arXiv:2410.09865*, 2024.
- [24] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [26] J. Huoa, W. Lia, Y. Shia, Y. Gaoa, and H. Yinb. Webcaricature: a benchmark for caricature face recognition. *arXiv preprint arXiv:1703.03230*, 2017.
- [27] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, Oct. 2024. 33 pages.
- [28] K. Karkkainen and J. Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, 2021.
- [29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
- [30] M. Kim, F. Liu, A. Jain, and X. Liu. DCFace: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023.
- [31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [33] G. Li, H. Hsu, C.-F. Chen, and R. Marculescu. Machine unlearning for image-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [34] R. Liu, B. Ma, W. Zhang, Z. Hu, C. Fan, T. Lv, Y. Ding, and X. Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2114–2123, 2024.
- [35] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale Celebfaces attributes (CelebA) dataset. *Retrieved August*, 15(2018):11, 2018.
- [36] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert. GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2023.
- [37] Y. Men, Y. Yao, M. Cui, Z. Lian, and X. Xie. DCT-Net: Domain-calibrated translation for portrait stylization. *ACM Transactions on Graphics (TOG)*, 41(4):1–9, 2022.
- [38] A. Mishra, S. N. Rai, A. Mishra, and C. Jawahar. IIIT-CFW: A benchmark database of cartoon faces in the wild. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part I 14*, pages 35–47. Springer, 2016.
- [39] S. Mishra, P. Shalu, and R. Singh. Enhancing face emotion recognition with facs-based synthetic dataset using deep learning models. In *International Conference on Computer Vision and Image Processing*, pages 523–531. Springer, 2023.
- [40] D. N. M. Nizam, D. N. Rudiyansah, N. M. Tuah, Z. H. A. Sani, and K. Sungkaew. Avatar design types and user engagement in digital educational games during evaluation phase. *International Journal of Electrical and Computer Engineering*, 12(6):6449, 2022.
- [41] K. L. Nowak and J. Fox. Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations. *Review of Communication Research*, 6:30–53, 2018.
- [42] OpenAI. ChatGPT: Conversational AI by OpenAI. OpenAI: <https://openai.com/chatgpt>, 2024.
- [43] OpenAI. DALL-E 3: OpenAI’s text-to-image model. OpenAI: <https://openai.com/dall-e>, 2024.
- [44] M. Oquab, T. Dariset, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [45] P. Panda, M. J. Nicholas, M. Gonzalez-Franco, K. Inkpen, E. Ofek, R. Cutler, K. Hinckley, and J. Lanier. AllTogether: Effect of avatars in mixed-modality conferencing environments. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pages 1–10, 2022.
- [46] I. Pavlov. Watermark detection, 2021. GitHub repository: <https://github.com/boomb0om/watermark-detection>.
- [47] I. Pikoulis, P. P. Filntisis, and P. Maragos. Photorealistic and identity-preserving image-based emotion manipulation with latent diffusion models. *arXiv preprint arXiv:2308.03183*, 2023.
- [48] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [49] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfelix, and F. Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 818–833, 2018.
- [50] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. SynFace: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021.
- [51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International*

- Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.
 - [53] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
 - [54] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
 - [55] A. Sauer, K. Schwarz, and A. Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.
 - [56] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. LAION-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
 - [57] K. Segaran, A. Z. Mohamad Ali, and T. W. Hoe. Does avatar design in educational games promote a positive emotional experience among learners? *E-learning and Digital Media*, 18(5):422–440, 2021.
 - [58] S. I. Serengil and A. Ozpinar. LightFace: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE, 2020.
 - [59] X. Shen, S. Lei, and J. Liu. Overview of cartoon face generation. In *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 6, pages 792–799. IEEE, 2023.
 - [60] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
 - [61] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.
 - [62] T. Stadler, B. Oprisanu, and C. Troncoso. Synthetic Data – Anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, Boston, MA, Aug. 2022. USENIX Association.
 - [63] S. Tripathy, J. Kannala, and E. Rahtu. ICface: Interpretable and controllable face reenactment using GANs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3385–3394, 2020.
 - [64] R. Webster, J. Rabin, L. Simon, and F. Jurie. This person (probably) exists. Identity membership attacks against GAN generated faces. *ArXiv*, abs/2107.06018, 2021.
 - [65] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3681–3691, 2021.
 - [66] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
 - [67] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2018.
 - [68] Y. Zhao, L. Yang, E. Pei, M. C. Ovemeke, M. Alioscha-Perez, L. Li, D. Jiang, and H. Sahli. Action unit driven facial expression synthesis from a single image with patch attentive GAN. In *Computer Graphics Forum*, volume 40, pages 47–61. Wiley Online Library, 2021.
 - [69] C. Zheng, M. Mendieta, and C. Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3146–3155, 2023.
 - [70] Y. Zheng, Y. Zhao, M. Ren, H. Yan, X. Lu, J. Liu, and J. Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2264–2272, 2020.