A Unified and Interpretable Emotion Representation and Expression Generation

Reni Paskaleva[3*], Mykyta Holubakha[1], Andela Ilic[2], Saman Motamed[1], Luc Van Gool[1,2], Danda Paudel[1]
[1]INSAIT, Sofia University, Bulgaria [2]ETH Zurich, Switzerland
[3]First Private Mathematical High School, Sofia, Bulgaria
firstname.lastname@insait.ai, anilic@student.ethz.ch

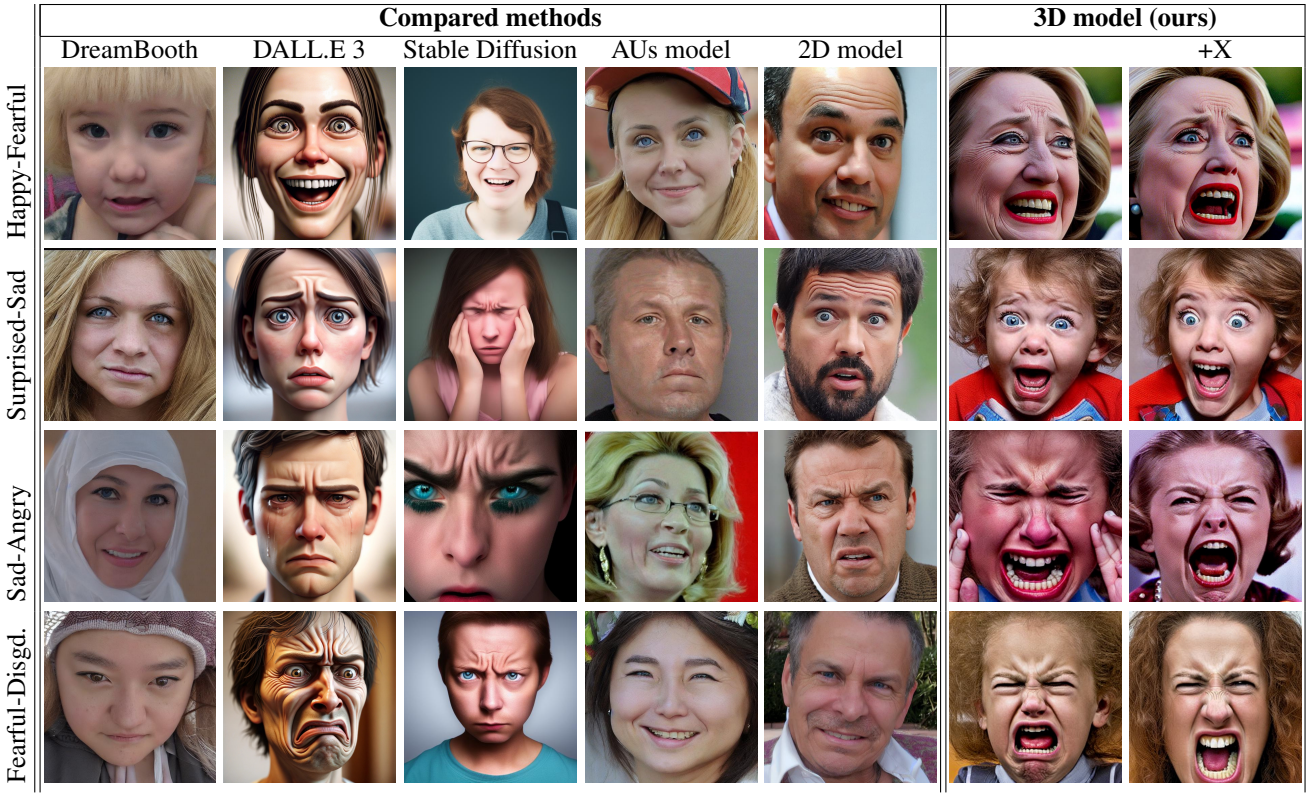

Figure 1. We show the capability of our continuous 3D-representation-based expression generation method in generating rich and compound expressions. An extra arbitrarily chosen expression component (+X) is added to the targeted compound on the left. The proposed 3D model performs the best compared to the 2D model and other competing methods. Our model shares the same settings with DreamBooth.

Abstract

Canonical emotions, such as happy, sad, and fearful, are easy to understand and annotate. However, emotions are often compound, e.g. happily surprised, and can be mapped to the action units (AUs) used for expressing emotions, and trivially to the canonical ones. Intuitively, emotions are continuous as represented by the arousal-valence (AV) model. An interpretable unification of these four modalities —namely, Canonical, Compound, AUs, and AV— is highly desirable, for a better representation and understanding of emotions. However, such unification remains to be unknown in the current literature. In this work, we propose an interpretable and unified emotion model, referred as C2A2*. We also develop a method that leverages labels of the non-unified models to annotate the novel unified one. Finally, we modify the text-conditional diffusion models to understand continuous numbers, which are then used to generate continuous expressions using our unified emotion model. Through quantitative and qualitative experiments, we show that our generated images are rich and capture subtle expressions. Our work allows a fine-grained generation of expressions in conjunction with other textual inputs and offers a new label space for emotions at the same time.*

Project & code: https://emotion-diffusion.github.io/
*This work was done as a part of INSAIT internship.

Category	AUs	Category	AUs
Happy	12,25	Sadly disgd.	4,10
Sad	4,15	Fearfully angry	4,20,25
Fearful	1,4,20,25	Fearfully surpd.	1,2,5,20,25
Angry	4,7,24	Fearfully disgd.	1,4,10,20,25
Surprised	1,2,25,26	Angrily surpd.	4,25,26
Disgusted	9,10,17	Disgd. surpd.	1,2,5,10
Happily sad	4,6,12,25	Happily fearful	1,2,12,25,26
Happily surpd.	1,2,12,25	Angrily disgd.	4,10,17
Happily disgd.	10,12,25	Awed	1,2,5,25
Sadly fearful	1,4,15,25	Appalled	4,9,10
Sadly angry	4,7,15	Hatred	4,7,10
Sadly surpd.	1,4,25,26	-	-

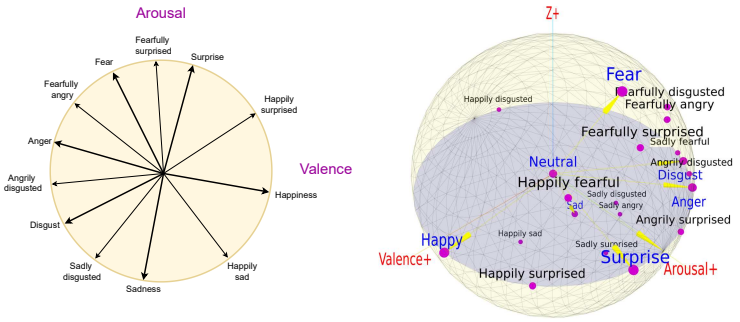

Figure 2. The compound emotion model on the left unifies the categorical emotions and the AUs based expressions [4]. The continuous emotion model of arousal-valence (middle) allows the mapping of some of the categorical emotion on the continuous space [39]. The proposed 3D-based emotion modelling largely unifies the both thereby allowing more combination of the compound emotions (right).

1. Introduction

Expressing emotions affects our lives by playing a vital role in day-to-day communications. We are interested in facial expressions, which are a primary means of such communication. Therefore, a generative model must generate realistic expressions for human-like communications. Human emotions and expressions however, are very complex even for humans to articulate with natural language. Among many, one possible reason is the used language for describing them – in particular the existing different modalities, which are even inconsistent with each other. This paper aims for a unified emotion model that is consistent and mappable to the existing ones, which makes our model also interpretable. The unified model is then used to enable generating rich facial expressions using text-to-image models, as shown in Figure 1. Further, we propose a method capable of understanding continuous expressions.

The most commonly used emotion models include basic Categorical [6], Compound [4], and arousal-valence (AV) [39] -based. The categorical model is simple, intuitive, and easy to annotate, while the compound emotion model is more complete. On the other hand, AV-based models are continuous where Categorical models can also be mapped, as shown in Figure 2 (middle). A popular physics-based modeling of expressions is Action Units (AUs) [30], which relies on the activation of the facial muscles. In fact, Compound emotions can be mapped to AUs, as shown in Figure 2 (left). We aim to map all Categorical, Compound, AV, and AUs in a common unified representation, as shown in Figure 2 (right), which we refer to as C2A2 (for Canonical, Compound, Action units, Arousal-valence). To the best of our knowledge, such unification is proposed for the first time in this paper. Our proposed unification offers a better representation, leading to more versatile emotion generation, in the context of this paper.

One major challenge of using a new emotion model is the missing associated labels. To address this problem, we first propose a 3D model such that it can exploit the existing 2D AV labels. Then, we propose a method to learn the additional third dimension without requiring any explicit supervision. In the language of basic categorical emotions, we lift up the "fear" towards the positive third-dimension, and the "sad" towards the negative side, as shown in Figure 2 (right). The choice of these two particular emotions is made to best cover the compound emotions presented in the Figure 2 (left). To learn the third dimension, our method leverages AUs-based modeling, where 3D vectors representing some compound emotions are first mapped to the action units followed by their supervision within a learning framework, inspired by GANmut [5]. This allows us to first generate the third-dimension labels (Z), which we later use to learn the conditional image generation.

On the image generation side, large text-to-image diffusion models [2, 28, 34, 42, 51] have emerged as a powerful way of generating high-quality images. However, the existing models cannot understand the continuous number required to represent the facial expressions of our interest. Therefore, we also develop a method that facilitates generating images conditioned upon text and a vector of continuous numbers that represents the target emotions of interest. More specifically, we propose to use a number encoder that maps the emotion-condition vector into the common text embedding space. The embedded numbers are then used together with the text embeddings to generate text and emotion-conditioned images.

We use the latent diffusion model [34] in the training setting of DreamBooth [37]. In this setting, we perform two parallel loss computations, one with and other without the emotion embedding. This regularizes the training and helps to preserve the knowledge of base diffusion model, thus allowing us to generate images with a rich expression and additional attributes described by the conditioning text input. Our experiments clearly demonstrate the superiority of the proposed 3D emotion model over the existing 2D AV, in the very same setup. Furthermore, our model that can

understand both text and numeral inputs provides very very convincing expression generation results.

Our major contributions can be summarized as:

- We propose an emotion model that unifies four different existing models in a common interpretable framework.
- We propose a method to annotate the emotion in the proposed emotion space by leveraging the AV and AUs.
- A number+text-to-image diffusion model is proposed to accommodate the proposed 3D numerical representation.
- Our results validate the 3D emotion model, annotation method, and number+text-to-image generation, by offering better quality and fine-grained control of expressions.

2. Related Work

Modeling human emotions is a century-long ongoing topic of study [7, 22, 38–41, 48]. A commonly used model is basic categorical emotions [6, 24, 36]. Other models are also used [12, 43]. Among these, the most commonly used are compound emotions [4], arousal-valence (AV) [39], and Action Units (AUs) [30]. The compound emotion representation also embeds the AUs, up to an extent, as shown in Figure 2. Similarly, AUs representation also embeds basic categorical and some compound emotions. However, to the best of our knowledge, there is no emotion model that offers better unification than the mentioned above.

Understanding and generating expressions are of high interest in computer vision [1, 3, 5, 8, 10, 11, 21]. Many generative works are utilized for the purpose of realistic manipulation of human emotions, including StarGAN [8], GANimation [10], SMIT [10], GANmut [5], ICface [20] and Neural Emotion Director [11]. These methods use generative adversarial networks (GANs) [15] to generate or manipulate the expressions expressed in the existing emotion models. Differently, a recent work [52] uses a diffusion model to generate landmark controlled 3D meshes for facial expressions. In this work, we introduce a new emotion model and propose a text-to-image diffusion method to generate images with expressions representing our targeted emotions. Nevertheless, we use the framework of GANmut to annotate images in our emotion representation space.

Diffusion-based methods [16, 19, 45–47] have become the go-to choice for image generation due to their synthesis quality and stable training. Recently, text-to-image diffusion methods [29, 32, 33, 35, 42] have shown promise in enabling an intuitive interface for users to control image generation, using natural language descriptions. However, fine-grained control and customized image generation has proven difficult with natural language descriptions alone [13, 25, 26, 49]. To address this problem, some existing works adapt pre-trained models to their targeted examples, either to find pseudo-words[14, 27] or fine-tune some parts of the pre-trained model [23, 37]. The pseudo-words are searched in the text embedding space of the text encoder (e.g, CLIP [31]). On the other hand, the fine-tuning methods such as DreamBooth [37] fine-tunes only the attention layers while preserving the generation capabilities of the original network. Therefore, we are interested in this setting and thus develop our method to augment DreamBooth [37] with fine-grain control over facial expressions.

Category	2D	3D	Category	2D	3D
Sadly disgd.	✓	✓	Happily disgd.	✗	✓
Fearfully angry	✓	✓	Sadly feraful	✗	✓
Fearfully surpd.	✓	✓	Sadly angry	✗	✓
Angrily disgd.	✓	✓	Fearfully disgd.	✗	✓
Happily surpd.	✓	✓	Angrily surpd.	✗	✓
Happily sad	✓	✓	Happily fearful	✗	✓
-	-	-	Sadly surpd.	✗	✓
Awed	-	- →	Happy+surprise+fear	✗	✓
Hatred	-	- →	Disgust+anger+fear	✗	✓
Appalled	-	- →	Disgust+surprise	✗	✗
Disgd. surpd.	✗	✗	-	-	-

Table 1. The compound emotions that can and cannot be represented by the proposed 3D representation of emotions. Our 3D model can represent 15/17 desired emotions (after mapping "Awed" and "Hatred" to composition of three basic emotions), whereas, 2D representation of AV can represent only 6/17.

3. The C2A2 Emotion Model

While aiming to generate compound and continuous expressions, we realized that the existing interpretable representations do not support our needs. Therefore, we proceed to modify the most suitable existing model, namely arousal-valence, as it already embeds the basic emotions in the 2D continuous space. In fact, this representation allows six compound emotions to be expressed, as shown in Table 1. In the same Table, it can be seen that other compound emotions of interest are not expressed using the 2-dimensional AV model. Therefore, we propose to represent the emotions in 3-dimensional space, while preserving the structure of the AV-based 2D model. More specifically, we lift the "fear" towards the positive third-dimension, and the "sad" towards the negative side, as shown in Figure 2 (right). This choice is made to best cover the most number of compound emotions, i.e. 15/17, as shown in the Table 1. In fact, we decompose two terminologies of [30], "Awed", "Hatred", and "Appalled", into the composition of basic emotions, happy+surprise+fear, disgust+anger+fear, and disgust+surprise, respectively. This leads to the compatibility of two additional compound emotions. Unfortunately, the emotion "Appalled" and "Disgustedly surprised" are not yet compatible with our emotion model. We choose to avoid modelling these two emotions to simplify our emotion model and make use of the AV labels. In fact, in our 3D representation, we use AV labels as the 2D coordinate and learn the third dimension, which we denote with a variable Z, using a method inspired by GANmut [5], with the help of the Action Units' labels, presented below.

3.1. Implicit Supervision for Z of C2A2

To learn the third dimension of C2A2, we extend the idea proposed in 2D linear GANmut model that has a conditional space parameterized with polar coordinates θ and ρ, which are both uniformly distributed $\theta \sim U([0, 2\pi])$, $\rho \sim U([0, 1])$. The angle θ indicates the category of the emotion, while the radius ρ tells more about its intensity. The expected behavior of the model is that the intensity increases with the distance from the center and the emotion transition between two basic emotions is smooth and continuous, reflecting the compound emotions in between.

Our method aims to learn a linear 3D model by fixing the positions of basic categorical emotions. We use the AffectNet [17] dataset. From the AV labels, the angles of the basic emotions are determined by averaging the AV labels corresponding to them. We fix the angles of basic emotions happiness, surprise, disgust, and anger vectors following the estimations. The extreme point of each of these emotions is set at the maximum distance from the center. On the other hand, fear and sadness, we lift above and under the AV plane respectively. The lifting is constrained in such a way that their projection to the AV plane corresponds to the expected 2D position, and makes 60° angle with the AV plane. Now, any emotion represented in our 3D space can be mapped back to AV by projecting on the XY-plane.

By lifting the two emotions to the third dimension, we have created a void in labels along the Z dimension. We wish to learn the labels along this dimension in an implicit manner, so as to avoid the need of direct annotations. Note from Table 1 that by making the modification we are able to represent nine additional compound emotions. More importantly, *these compound emotions can be mapped to the AUs, which are also continuous in nature* (please, refer the left side of Figure 2). Now, our interest is to exploit the continuous labels of AUs, which can conversely help us in the 3D space C2A2. In this work, we first design the mapping process for AUs, and then exploit them to learn the 3D space of C2A2 by using the conditional space learning framework, from weak labels, of GANmut [5].

Learning the conditional space is based on conditioning the samples from mini batches that correspond to the basic categorical emotions, but this time we include also the compound ones. While learning Z, we supervise our model also by AV labels, say v_{va}. Therefore, we add the following new loss to the discriminator's objective function,

$$\mathcal{L}_{av} = \mathbb{E}_{x,z}[||D_{coor}(x) - v_{av}||_2^2], \tag{1}$$

where, $D(.)$ is the discriminator network, and x is the input image. The second loss added to the GANmut discriminator's objective function is the AU loss. Since AffectNet does not provide action units' labels, they were manually mapped from the valence and arousal values. Therefore, labels of the real images are limited to 12 possible sets of AUs, which could be found in the AV plane (please, refer to the Figure 2 (left)). The mapping starts by dividing the space between each basic and surrounding compound emotions into two parts, such that one half could be still considered the basic emotion, while the other half goes into the part covered by the compound one.

$$Y \longrightarrow \text{Emotion category} \longrightarrow AU \overset{\mathcal{L}_{AU_Y}}{\longleftrightarrow} \hat{AU}_{real}$$

We obtain the pseudo-labels $\hat{AU}_{real}$ using the OpenGraphAU tool [9]. The OpenGraphAU provide us the activation probability of all 41 actions units. Based on Figure 2 (left), we need only 15 action units in total to decide on the compound emotions of our interests. The pseudo-labels of AUs are used to compute an additional loss between the labeled activation probabilities and the estimated ones,

$$\mathcal{L}_{AU_Y} = \mathbb{E}_{x,Y}[KL(D_{au}||AU_Y) + KL(AU_Y||D_{au})]. \tag{2}$$

Here, $KL(.||.)$ is the Kullback-Leibler Divergence, and $Y = [A, V, Z]^\intercal$ is the 3D conditional vector. During training, the generator is conditioned on two mentioned mini batches and sampling was performed along the basic and compound emotion vectors (or in their proximity).

3.2. Unprojecting Images along Z

Once the conditional space is learned using the GANmut framework, we obtain the labels for Z in rather a straightforward manner. Although the conditional space is implicitly learned by the earlier training, the images still need to be mapped to the conditional space to obtain the Z labels. This could possibly be done more accurately by using techniques reported [50]. However, we use a simple approach and obtain the sought labels directly from the discriminatory network. Let $\hat{Z} = D(x)$ be the z-dimensional label predicted by the discriminator for a given image x, then we use $Y = [A, V, \hat{Z}]$ as the emotion label corresponding to that image in the proposed representation.

4. Hybrid Text-to-Image Generation

Describing compound emotions using natural language descriptions does not always lead to faithful representation of the intended emotion (see Figure 1) in text-to-image models. To introduce granular control over such emotions, while taking advantage of large text-to-image diffusion models, we use a number encoder that encodes the 3D emotion vector $Y \in \mathbb{R}^3$ such that, when concatenated with the encoded text description, it can depict the described face with the intended emotion represented by the 3D vector.

During training, the prior loss $\mathcal{L}_{prior}$ is tasked with mitigating language drift and overfitting on training images while the reconstruction loss $\mathcal{L}_{recon}$ enables the desired control of the generated faces' expressions. Figure 3 shows

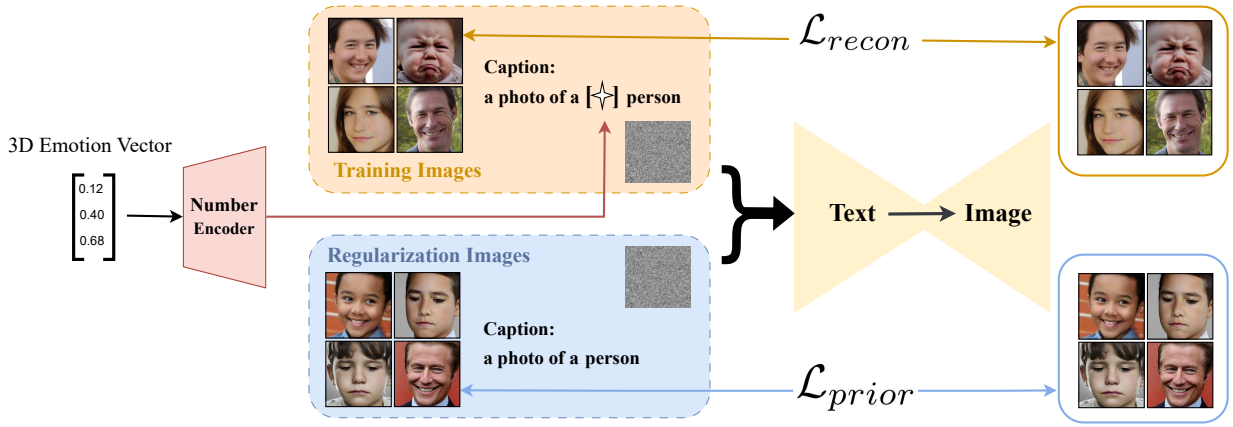

Figure 3. We use a number encoder that embeds the continuous 3D representations of emotions. The embedded numbers are fused with the text embedding before decoding into number+text-to-image generation. The learning is done using the frozen text-encoder and shared image decoder. During learning, our method uses prior preservation and emotion reconstruction loss, similar to DreamBooth [37].

Figure 4. Top three rows: images sampled around a circle (angle of AV on top) at different learned Z of our 3D model. Bottom: the same circle for 2D model. Not that the 3D model is clearly superior than 2D. The images on first and last rows may directly be compared.

how C2A2 effectively augments text-to-image models with a number encoder. We train our method on AffectNet dataset with and without 3D labels Y, which are obtained using the method described in the previous section.

Number Encoder. We use an MLP-based number encoder $E_y = \phi_\theta(Y)$ that projects the emotion vector $Y \in \mathbb{R}^3$ to a higher dimensional embedding. Note that the DreamBooth [37] framework uses CLIP [31] as the text encoder, following this we map our emotion vector to the embedding $E_y \in \mathbb{R}^{768}$. This embedding is merged with the embeddings from the text encoder to condition the image decoder.

Learning number+text-to-image generation. To enable the joint condition of the numbers and text, we first map the emotion vectors to a higher dimensional embedding. These embeddings are then used together with the text embeddings, followed by the text-to-image generator training using $\mathcal{L}_{prior}$ and $\mathcal{L}_{recon}$ losses, similarly as in [37]. During the inference, text and number embeddings are fused to generate images with continuous emotions and given text description. Please, refer Figure 7 for such examples.

5. Experiments

Datatest. The training of our models and baselines utilize AffectNet [17], recognized as the most extensive dataset in affect computing, comprising $\approx$ 1M images sourced from the Internet. Searches on prominent search engines were conducted using 1250 keywords linked to emotions across six languages. Remarkably, 450K of these images received manual annotations from 12 specialists, categorizing them into basic emotions and AV labels. Given our objective to generate intricate and nuanced emotions in a highly varied context, this dataset emerged as the perfect selection.

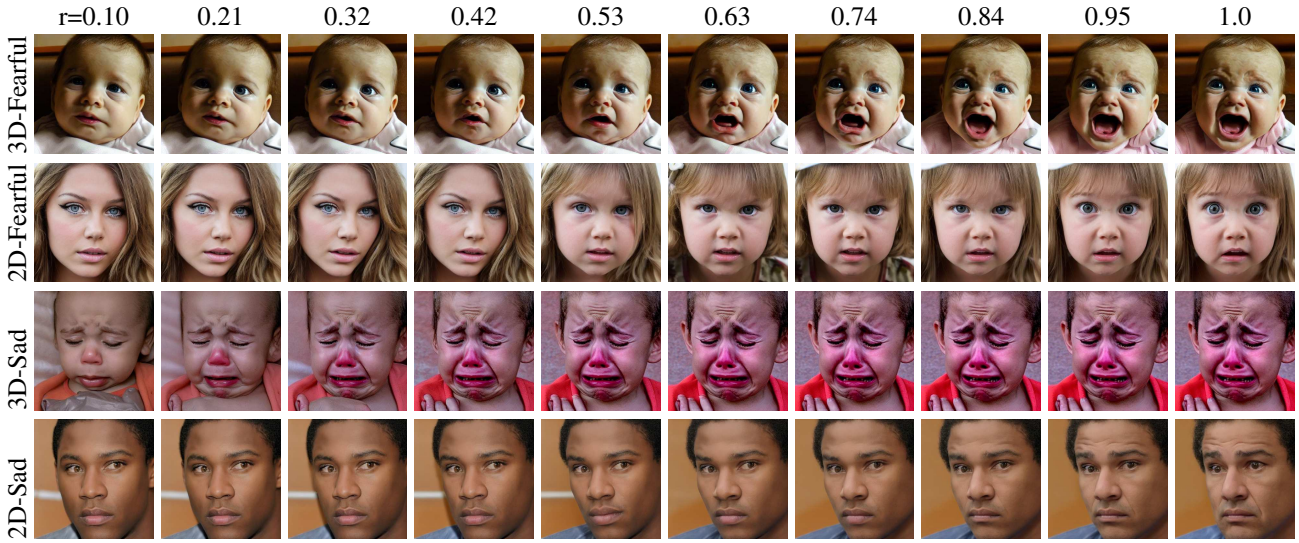

Figure 5. Both our 2D and 3D methods understand the emotions represented as continuous numbers. For 3D model, we showcase the behaviour towards the learned Z. These images illustrate that our learned representation is indeed continuous. Better viewed zoomed in.

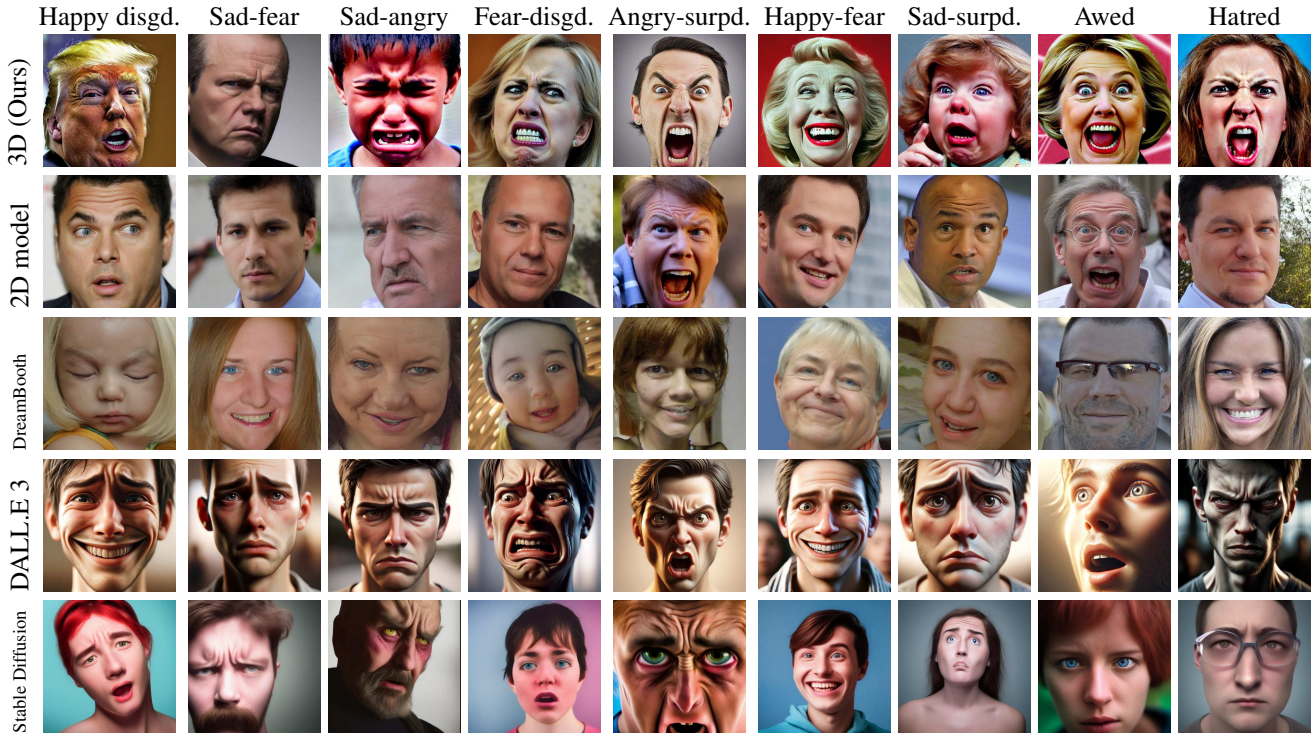

Figure 6. Nine compound emotions that can be represented by our 3-dimensional C2A2 but cannot be represented by 2-dimensional AV model (please refer Table 1 to associated the emotions (top, here) to the 3D vs. 2D representations). When compared with other methods, our 3D-based representation is clearly superior, thanks to its richer representation and the continuous number understanding capabilities.

Implementation details. We implemented two parts proposed in this paper in two different settings. The first part of our implementation is based on GANmut [5]. We set the number of training iterations to 1M. We adopted the same training strategy of GANmut, and the same hyper-parameters. For the second part, we implemented a Dreambooth-like approach on top of the Stable Diffusion model. We performed 2 experiments where we generated emotions from the 2D and the 3D spaces. We also tried several variation of text+number inputs, with limited success. That is why we use both a text and a number embeddings. During the training and the generation process we

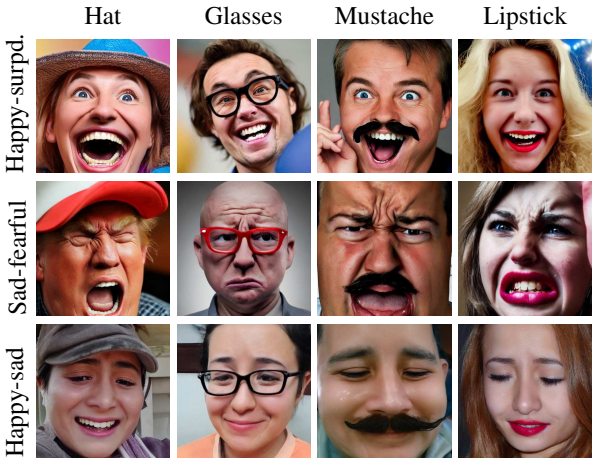

Figure 7. Our method can preserve the attributes from the base network. This allows us to perform meaningful number+text-to-image generation. Expression for Compound emotions (left) represented in numbers are generated with text attributes (top).

used "sks" as a placeholder token which is used to identify a specific emotion. We experimented with different text prompts such as "a photo of a sks person" and "a colourful photo of a sks person", and came to the conclusion that the best performing one is "a high-quality realistic color photo of a sks person", so we used it as a text prompt in all of our experiments. For regularization we used 100K randomly chosen images from the training datasets with the same prompt and all of the coordinates corresponding to their emotions were set to 0.

Evaluation metric. Similar to GANmut, we employ the modified Fréchet Inception Distance [18], termed as Fréchet Emotion Distance (FED), for assessing emotions. For calculating FED, we trained VGGNet [44] on AffectNet for emotion classification. This involves inputting real and generated images into VGGNet and extracting features proximal to the ultimate classifier. We then assume Gaussian distributions for both feature sets and compute their Fréchet distance. Our goal is to minimize the FED value. To determine FED, we uniformly randomly sample images in every instance. More evaluation methods are also presented, detailed in their respective subsections. For approximating human emotion assessment, we utilize the softmax score from the trained VGGNet. Additionally, we also engage 8 psychologists for evaluation through an expert user study.

Baselines. We use DreamBooth [37] trained on the AffectNet as our baseline for both quantitative and qualitative evaluations. Additionally, we use recent and popular image generation methods for further comparisons. More importantly, we perform exhaustive comparisons of our method that uses 2D AV representation against the proposed 3D representation. Please, refer to our supplementary material for more details, further analysis, and visualizations.

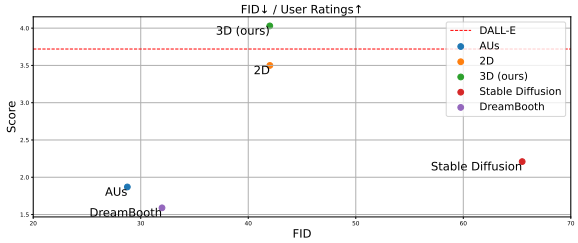

Figure 8. FID and user ratings (Scores) for six different methods. Our 3D representation based method presents the best combination between low FID and high user rating score. Two methods with low FID often fail to represent the targeted expressions.

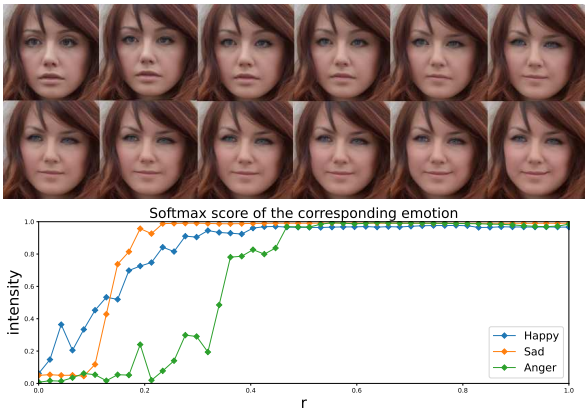

Figure 9. Top: subtle control of expression between r=0.02 and r=0.3 along the "happy" axis. Bottom: softmax scores with increasing radius away from the neutral to 3 canonical emotions.

5.1. Qualitative Results

We present five sets of qualitative results, as depicted in Figures 4, 5, 6, 7, and 9 demonstrating the efficacy of our methods. Figure 5 reveals that both our 2D and 3D models effectively interpret emotions represented as continuous numbers, with the 3D model exhibiting behavior towards the learned Z, indicating a continuous learned representation. Specifically, the top three rows of Figure 5 display images sampled around a circle at different learned Z values of our 3D model, with AV values annotated above. The bottom part of the same figure contrasts this with the 2D model, using the same circle in the AV space to generate expressions. Notably, the 3D model's results are significantly superior to those of the 2D model, as they can be directly compared in the first and last rows.

Figure 6 focuses on nine compound emotions that our 3D C2A2 model can represent but are beyond the scope of the 2D-AV model. This comparison, detailed in Table 1, underscores the superiority of our 3D-based representation, attributing to its richer representation and understanding of continuous numbers. For comparison, we also provide results obtained using Stable Diffusion [34], DreamBooth [37], and closed-source popular DALL-E 3. Figure 7 illustrates how our method preserves attributes

Method	FED (↓)	ERE (↓)	SS (↓)
DreamBooth [37]	61.147	–	–
2D-AV model	20.347	0.0774	0.905
3D model (Ours)	**16.060**	**0.0536**	**0.806**

Table 3. Fréchet Emotion Distance (FED), Emotion Reconstruction Error (ERE), and Smoothness Score (SS) for three different method. Our 3-dimensional representation of emotions clear outperform the other alternatives, when trained on the same settings.

from the base network, enabling meaningful number+text-to-image generation. Here, expressions for compound emotions, represented as numbers, are generated alongside text attributes, showcasing the versatility and robustness of our approach in handling complex emotional representations. Lastly, Figure 9 illustrates an example of subtle control.

5.2. Quantitative Results

We conduct two sets of quantitative evaluations. First we provide the results using the metric from GANmut [5]. Later, the evaluation by the human psychologists is provided, along with FID, confirming the significance of our 3D representation model, by a large margin, in generating meaningful images. The user rating and FID trade-off is summarized in Figure 8 for six different methods.

FED, ERE, and Smoothness. Together with the FED, we use the Emotion Reconstruction Error (ERE) and the Smoothness Score (SS). To calculate FED, we sample 50K images and compute the FID using the emotion features. For ERE, we conduct a uniform search for target emotion, using a sample budget of 500 images. Multiple runs are performed for each of seven basic emotions. Then ERE is the averaged emotion reconstruction errors between target and closest images. The SS is determined by using the VGGNet classifier with increasing emotion intensity. We follow [5] for these metrics, where readers can also find more details.

The FED, ERE, and SS, obtained by DreamBooth [37], 2-dimensional AV based representation, and the proposed 3-dimensional representation are reported in Table 3. Note that we train DreamBooth on AffectNet to obtain the reported results, which does not use 2D or 3D emotion labels. These results again highlighting the superiority of the our 3-dimensional C2A2 emotion model and method developed to perform number+text-to-image generation.

Expert user (psychologist) study. We asked 8 expert psychologists to rate 315 images (per method) on the scale from 1 to 5 (5 being the highest) by how much they agree with the coordinates of the images. We also provided them with the nearest emotion to every image. More details of our study is in the supplementary materials. The obtained results are reported in Table 2, which shows that the experts clearly prefer the images generated by the proposed 3D representation, for individual compound emotions as well as overall emotions. This further validates our emotion representation and expression generation methods.

6. Conclusion

We proposed a novel, unified and interpretable emotion representation that is capable of expressing nine additional compound emotions from the continuous 3-dimensional space. The continuous and 3D aspect of our representation allowed us to generate images with multitude of expressions. To facilitate such generation, we proposed two methods; one to recover the additional emotion axis Z, and another to generate images using the continuous vectors representing emotions in a DreamBooth-like setting. Both our qualitative and quantitative results showcase the superiority of the proposed emotion representation and the method for number+text-to-image generation. Furthermore, we showcase the capability of our method in generating compound expressions together with other facial attributes. We seek to extend our method along the temporal dimension by learning from the video examples as our future work.

Limitations and ethical statement. The emotion conditioning in our model is not entirely disentangled. This is particularly evident with attributes akin to age, shown in Figure 5. Our method in processing text descriptions beyond facial attributes needs further exploration. In this work, we controlled the expression without preserving the identity of the person. A future extension could benefit from adding identity preserving properties to our model. Central to our research ethos is a commitment to ethical and responsible data usage, with a strong focus on fostering socially responsible applications.

Acknowledgements. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

Model	Happy-disgd.	Sad-Fearful	Sad-Angry	Fearful-Disgd.	Angry-Surpd.	Happy-Fearful	Sad-Surpd.	Average	Average Overall
3D model (Ours)	**3.51**	**4.06**	**4.18**	**4.52**	**4.74**	**4.32**	**4.02**	**4.19**	**4.03**
2D-AV model	2.71	3.00	2.67	3.09	3.67	3.41	2.50	3.01	3.50
AUs model	1.17	1.55	1.77	1.83	1.60	1.69	1.92	1.76	1.87
Stable Diffusion	1.32	2.15	2.31	2.31	1.89	2.50	2.07	2.04	2.21
DALL.E 3	2.25	3.43	3.45	3.36	3.94	3.04	2.60	3.31	3.72
DreamBooth	1.94	1.32	1.07	1.33	1.23	1.14	1.59	1.47	1.59

Table 2. Average ratings (out of 1-5) for 7 compound emotions that can be represented by our 3D C2A2 but cannot be represented by the 2D-AV model, the average of these 7 emotions (second last) and the average among all emotions (last column) used in our study.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 367–374, 2018. 3

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022. 2

[3] Andrew J Calder and Andrew W Young. Understanding the recognition of facial identity and facial expression. *Facial Expression Recognition*, pages 41–64, 2016. 3

[4] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 2, 3

[5] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, Luc 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. 2, 3, 4, 6, 8

[6] P Ekman and W Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 1971. 2, 3

[7] W. V. Ekman, P.and Friesen. The facial actin coding system: a technique for the measurements of facial movements. 1978. 3

[8] Choi, Yunjey, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE conference on computer vision and pattern recognition.*, 2018.. 3

[9] Luo, Cheng, et al. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022.. 4

[10] Pumarola, Albert, et al. Ganimation: Anatomically-aware facial animation from a single image. *Proceedings of the European conference on computer vision (ECCV).*, 2018.. 3

[11] Papantoniou, Foivos Paraperas, et al. Neural emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022.. 3

[12] Vielzeuf, Valentin, et al. The many moods of emotion. *arXiv preprint arXiv:1810.13197*, 2018.. 3

[13] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 3

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 3

[17] Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing 10.1*, pages 18–31., 2017. 4, 5

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017. 7

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[20] Tripathy, Soumya, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. *Proceedings of the IEEE/CVF winter conference on applications of computer vision.*, 2020. 3

[21] Daeha Kim and Byung Cheol Song. Emotion-aware multi-view contrastive learning for facial emotion recognition. In *European Conference on Computer Vision*, pages 178–195. Springer, 2022. 3

[22] Robert E. Kraut and Robert E. Johnston. Social and emotional messages of smiling: An ethological approach. 1979. 3

[23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 3

[24] Randy J Larsen and Edward Diener. Promises and problems with the circumplex model of emotion. 1992. 3

[25] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[26] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 3

[27] Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert concepts beyond object appearance in text-to-image diffusion models, 2023. 3

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation

and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[30] Ekman, Paul, and Wallace V. Friesen. Facial action coding system. 1978. 2, 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[34] 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. 2, 7

[35] 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. 3

[36] Barbara H Rosenwein. Problems and methods in the history of emotions. *Passions in context*, 1(1):1–32, 2010. 3

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 5, 7, 8

[38] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980. 3

[39] James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 1977. 2, 3

[40] James Russell, Maria Lewicka, and Toomas Niit. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57:848–856, 1989.

[41] James Russell, Jo-Anne Bachorowski, and José-Miguel Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54:329–349, 2003. 3

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[43] Klaus R Scherer et al. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000. 3

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 2014. 7

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[48] Lilu Tang, Peijun Yuan, and Dan Zhang. Emotional experience during human-computer interaction: A survey. *International Journal of Human–Computer Interaction*, pages 1–11, 2023. 3

[49] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1910, 2023. 3

[50] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022. 4

[51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[52] Kaifeng Zou, Sylvain Faisan, Boyang Yu, Sébastien Valette, and Hyewon Seo. 4d facial expression diffusion model. *arXiv preprint arXiv:2303.16611*, 2023. 3