EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models (Supplementary Material)

Jingyuan Yang, Jiawei Feng, Hui Huang*
Shenzhen University
{jingyuanyang.jyy, fengjiawei0909, hhzhiyan}@gmail.com

1. Construction Details

Here we present the detailed network design in our method.

1.1. Dataset

Our dataset is derived from the large-scale EmoSet. There are several details worth mentioning. Considering the aim of Emotional Image Content Generation (EICG), we focus on the image content. Therefore, when deriving our training dataset, we only choose images with scene or object attributes. If an image contains both scene and object labels, scene attributes will be preserved to serve as our semantic guidance. If an image consists multiple object labels, the first object with the highest confidence score will be restored. After implementing the filtering strategies, we can successfully derive a subset of EmoSet, where number of images is decreased to 75,460. Additionally, We find that imbalanced emotions may lead to sub-optimal generation results. Thus we employ random oversampling by duplicating images in categories with fewer samples to align with the larger ones.

1.2. Emotion Space

The aim for emotion space is to find powerful emotional representations where similar emotions are gathered together and dissimilar ones are set apart. Intuitively, we employ layers before the last fully connected layer in the classifier as our emotion encoder. Since the emotion classifier can achieve high accuracy, we assume that the emotion encoder can well capture the emotional relationships.

To construct the emotion space, we utilize the entire EmoSet and employ ResNet50 as backbone to train an eight-class emotion classifier. Specifically, the last layer is replaced with a fully connected layer of size (2048,768) with an activation function, a dropout layer, and a fully connected layer of size (768,8). The emotion encoder φ, capable of capturing emotion representation, is constructed by removing the last layer of the trained emotion classifier. We learn the emotional representations from all the 118,102 samples in EmoSet. As eight emotions in the emotion space are separated, we simulate their distributions with Gaussian function and construct it by calculating the mean and standard deviation of each emotion. By calculating the similarity between the real distribution and simulated one, we find that Gaussian function is suitable for our task.

*Corresponding author

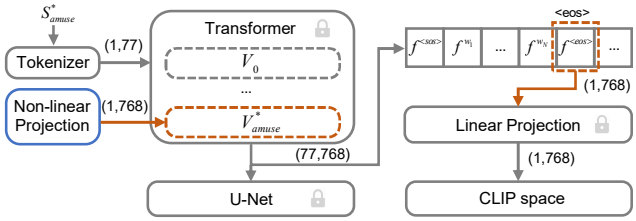

Figure 1. Detailed illustration of the mapping network.

1.3. Mapping Network

To bridge the gap between emotion space and CLIP space, we introduce a mapping network with a non-linear projection, followed by a transformer and linear projection. We attempt to use different mapping strategies by altering the number of layers, as shown in Table 1. We finally draw a conclusion that a two-layer MLP is the best choice with a perfect non-linear intensity.

In the mapping network, we employ a non-linear projection consisting of two fully connected layers with a ReLU activation function sandwiched in between. The FC layers are designed with the size of (768,1024) and (1024,768). The parameters inside the transformer and linear projection are taken from the same version of CLIP text encoder inside the stable diffusion. Specifically, emotion features are mapped through a non-linear projection into the embedding layer of the transformer.

Image Encoder and U-Net leverage parameters sourced from vae and U-Net within runwayml/stable-diffusion-v1-5. Figure 1 shows the specific design of the mapping network. Following [1], we opt the end-token embedding $f^{<eos>}$ with dimension (1, 768), followed by a linear projection into the CLIP space.

Table 1. Ablation study on the layer number of the non-linear projection, involving five metrics.

Method	FID ↓	LPIPS ↑	Emo-A ↑	Sem-C ↑	Sem-D ↑
DB(5 images)	160.46	0.638	71.03%	0.530	0.0129
DB(5000 images)	60.87	0.652	50.77%	0.509	0.0163
DB(balance strategy)	54.15	0.648	55.38%	0.490	0.0146
DB	46.89	0.661	70.50%	0.614	0.0178
SD(caption finetune)	54.81	0.697	72.82%	0.584	0.0219
Ours	**41.60**	**0.717**	**76.25%**	**0.633**	**0.0335**

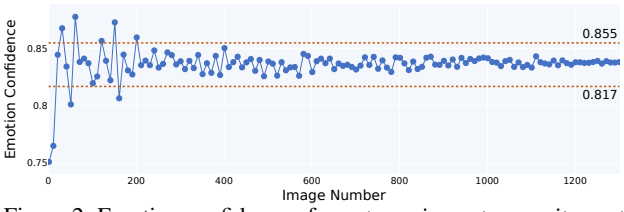

Figure 2. Emotion confidence of *sports equipment* on *excitement*.

1.4. Attribute Loss

Since attribute loss needs the guidance of object/scene, all images in the derived subset can be utilized in its optimization process. As mentioned above, we preserve only one semantic label for each image to compute the attribute loss, considering that each feature in emotion space can only has a unique optimization direction. That is, if a sample point is simultaneously optimized to the direction of *amusement park*, *trees*, *people*, and *balloons*, the network will learn attributes in a chaotic mode.

1.5. Emotion Confidence

Considering the fact that EmoSet comprises 118,102 emotional images collected from different sources, *i.e.*, openverse, pexels, pixabay and rawpixels, we assume that EmoSet can represent the visual emotion distribution in real world correctly and comprehensively.

We further train an emotion classifier on the CLIP space, which achieves an accuracy of 83%. Regarding to the high value in emotion accuracy, we believe that the pre-trained emotion classifier can separate different emotions effectively. Thus, we use this classifier to calculate the emotion confidences between emotions and attributes, hoping to filter emotional attributes from the emotion-agnostic ones.

Compared with real-world data, Calculating emotion confidence with EmoSet may inevitably introduce some biases. However, the bias in Figure 2 is tolerable. When image number reaches 200, emotion confidence falls within a 95% confidence interval of [0.817, 0.855], as calculated by Bootstrap. Experiments have also verified the effectiveness of emotion confidence both quantitatively and qualitatively.

2. Implementation Details

Our experiments are implemented based on PyTorch and performed on eight Quadro RTX 6000 with 24GB memory. We use the pre-trained stable-diffusion-v1-5 model and clip-vit-large-patch14 model. To train the non-linear projection, we utilize the subset of EmoSet. Emotion space is constructed with an emotion extractor trained by the entire EmoSet. Specifically, the dataset was split into 80% for training, 5% for validation, and 15% for testing, following the previous work. Emotion loss in the first stage is trained with a learning rate of 0.001 and a batch size of 32 to optimize while LDM loss and attribute loss in the second stage is trained with a learning rate of 0.001 and a batch size of 1. For the user study, we hired 14 healthy Asian volunteers comprising 10 males and 4 females, with ages ranging from 22 to 56.

2.1. Comparative Methods

In the comparative methods (*i.e.*, SD, TI, DB), stable diffusion is sourced from runwayml/stable-diffusion-v1-5 and training data encompasses all the images in the EmoSet subset, aligning with the proposed method. We adhere the same training steps and learning rate with their original settings. Unlike the customized image generation task, multiple concepts may coexist in one emotion category, rendering the use of the original 3-5 input images impractical. Take DB as an example, we present three additional settings in Table 1, involving different training sizes (*i.e.*, 5, 5000, all) and strategies (*i.e.*, w, w/o attribute balancing), where results in our paper achieve the best performance across five metrics. Notably, 5-image setting yields better result on Emo-A but exhibits a large gap on FID and Sem-D, which arises from a straightforward duplication of the learned 5 images.

Besides, we try to finetune with all captions in EmoSet. A primary challenge in EICG is the scarcity of manually annotated data. Another critical issue is the *affective gap*, for which we build a mapping network to interprete abstract emotions with concrete semantics. Since EmoSet is not annotated with captions, we employ BLIP2 to assign each image a text description by utilizing Salesforce/blip2-opt-2.7b. We then use captions with emotion labels to fine-tune SD with the whole EmoSet and report the result in Table 1, which achieves comparable results on five metrics.

3. Evaluation Metrics

As we introduce a new task, namely EICG, specially-crafted metrics should be designed in purpose. Rather than the commonly-used FID and LPIPS, we propose custom metrics to estimate the emotion fidelity, semantic clarity and semantic diversity of the generated images. The details on each evaluation metric are described as follows:

FID We adapt the FID to quantify the distribution distance between generated and real image. The lower the score is, the better quality the generation process achieves. In particular, we generate 1,000 images for each emotion and then calculated the FID score across all of them against our derived subset.

LPIPS To assess the overall image diversity, we employ LPIPS, *i.e.*, Learned Perceptual Image Patch Similarity. We randomly select P_i pairs of images for each emotion and calculate the LPIPS score by averaging the distances between these pairs. Finally, we average the LPIPS scores for each emotion to obtain the overall LPIPS:

$$LPIPS = \frac{1}{C}\frac{1}{P_i}\sum_{i=1}^{C}\sum_{p=1}^{P_i} LPIPS(a_p^i, b_p^i), \quad (1)$$

where C represents the total number of emotions, P_i denotes the number of sample pairs in emotion i and $LPIPS(\cdot)$ represents the LPIPS score between the p-th pair of image a and image b in emotion i.

Emo-A Emo-A is devised as a metric to assess emotion faithfulness. We utilize the pre-trained emotion classifier to predict the emotion of the generated images, and compare it with the targeted emotion, where only the correctly generated ones contribute to the final accuracy.

Sem-C We employ the pre-trained object classifier from ImageNet and the scene classifier from PLACES365 to classify the N images generated for each emotion. We take the highest probability between these two classifiers to construct the semantic clarity:

$$Sem\text{-}C = \frac{1}{N}\sum_{n=1}^{N} \max(v_{obejct}(x_n), v_{scene}(x_n)), \quad (2)$$

where N represents the total number of generated images, v_{obejct} and v_{scene} denote the classifier of object and scene.

Sem-D We randomly sample P_i pairs of images for each emotion and calculate the Mean Squared Error (MSE) between their CLIP image embeddings as the Sem-D score for that emotion. Then, we calculate the averaged Sem-D scores for all eight emotions:

$$Sem\text{-}D = \frac{1}{C}\frac{1}{P_i}\sum_{i=1}^{C}\sum_{p=1}^{P_i} MSE(\tau_\theta(a_p^i), \tau_\theta(b_p^i)), \quad (3)$$

where C represents the number of emotion, P_i denotes the number of samples pair in emotion i, MSE implies the mean squared error, τ_θ indicates the CLIP text encoder.

4. Additional Results

In Figure 4 and Figure 5, we present the comprehensive results on comparisons with the state-of-the-art methods and ablation studies, where Figure 4 contains positive emotions including *amusement*, *awe*, *contentment* and *excitement* while Figure 5 contains negative emotions including *anger*, *disgust*, *fear* and *sadness*. In the main paper, we have already presented *awe*, *anger* and *contentment* and here we add the remaining five emotions for completeness.

In Figure 4, we observe that the comparison methods are mainly concentrated on textures and colors. For example, in *amusement*, they can capture some emotion elements such as a smiling face and the Ferris wheel in amusement park, but they can hardly present them in a correct structure. In *excitement*, these methods can learn excited people but fail to place them in the accurate places. We can conclude that it may caused by the lack of semantic guidance, where our method can well-capture not only emotional elements in low-level but also high-level semantic relationships. In Figure 5, surprisingly, we find that even though compared methods can not generate images with clear semantics, they are still able to express *disgust* and *fear* to some extent. Considering these emotions are special for their nothingness and emptiness, where explicit contents are not that necessary. However, our method still evokes the corresponding emotions at the highest intensities. For the ablation study, we notice that the images distorted terribly with LDM loss alone, indicating the significance of attribute loss. Without emotion confidence, our method is prone to learn emotion-agnostic semantics caused by the unbalanced data distribution, where plants and trees appear mostly.

In Figure 6, we present the emotion decomposition results on all eight emotions, where each of them is represented with the six most correlated semantics. When thinking of *awe*, people tend to imagine *mountain snowy*, *flower field*, *valley*, *canal*, *waterfall* and *gulf*, where beautiful landscapes always bring us the feeling of respect and shock. For *anger*, we may easily relate it to *poster*, *tiger*, *gun*, *tank*, *helmet* and *army base*, which are concepts closely related to war. The correlations between emotions and contents in Figure 6 are gradually formed during the process of human evolution, where people easily experience certain emotions when viewing such objects or being within such scenes. Notably, the machine decomposed results are also highly aligned with human cognition, suggesting our method is not only effective in generating emotional images but also interpretable for human viewers.

We further combine emotions with neutral objects to create some interesting and meaningful emotional creations in Figure 7. In *amusement*, emotional elements are mostly *amusement park*, *smiling face*, *princes dress*, *balloon* and *light show*. Regarding to *awe*, emotional elements are always *mountain snowy*, *blue sky*, *ocean*, *wedding* and *tower*.

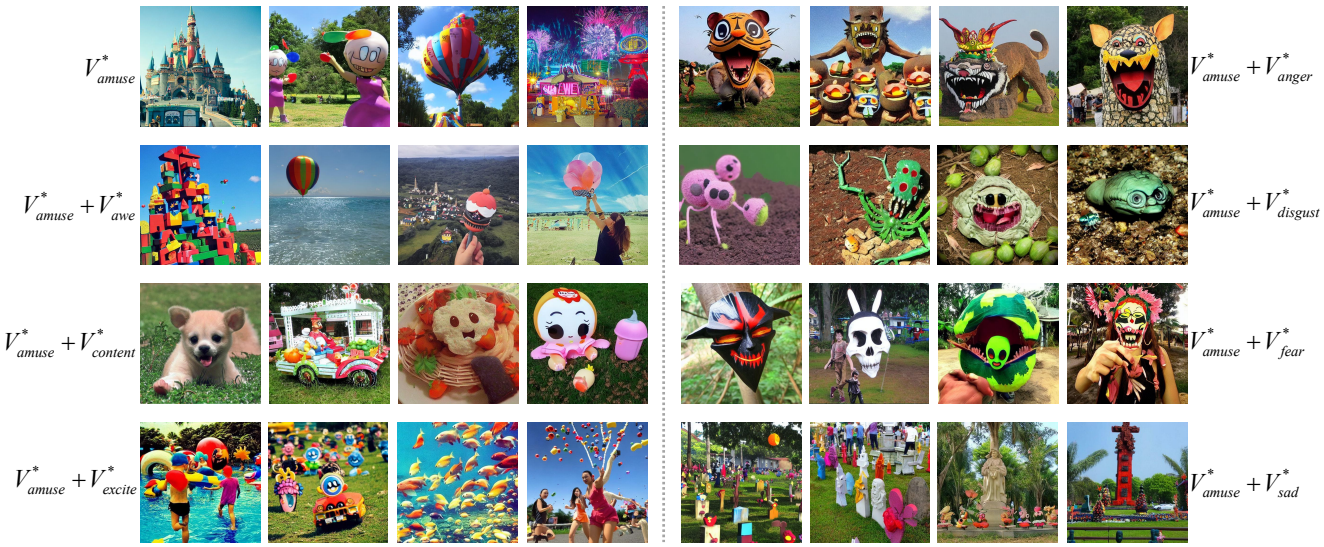

Figure 3. Emotion fusion, where we fuse the learned *amusement* embedding together with other seven emotions respectively.

Considering *contentment*, emotional elements are *flower* and *bedroom* while *excitement* are represented by *football* and *athletic field*. For *anger*, there comes *tiger*, *red monster* and *fire* while *disgust* corresponds to *rubbish* and *dirt*. While *fear* is pointed to *scared mask* while *spider* and *skull*, *cemetery* and *sculpture* can well represent *sadness*. With such elements, one can generate images with emotional creations automatically.

Eventually, we fuse different emotions together, hoping to explore more in emotion creation. In main paper, we only present the combination of *amusement* and *awe*, as well as *amusement* and *fear*. In Figure 3, we expand the results to other seven emotion and have some interesting findings. For example, when combining *amusement* and *contentment* together, there are *funny dog*, *colorful car*, *cute dishes* and *pleasant toy*, which can be seen as a success combination of these two emotional contents. While *anger* mostly concentrate on *furious but funny toy tigers*, *sadness* create some *colorful cemeteries*. These initial results are interesting and promising, which may be useful for emotional art creation. From the above applications we can witness the potential and significance of EICG.

References

[1] 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, 2021. 1

Figure 4. Comparison with the state-of-the-art methods and ablation studies of our method on four *positive* emotions.

Figure 5. Comparison with the state-of-the-art methods and ablation studies of our method on four *negative* emotions.

Figure 6. Emotion decomposition, where each of them can be decomposed to a set of emotional concepts. We list the most correlated ones.

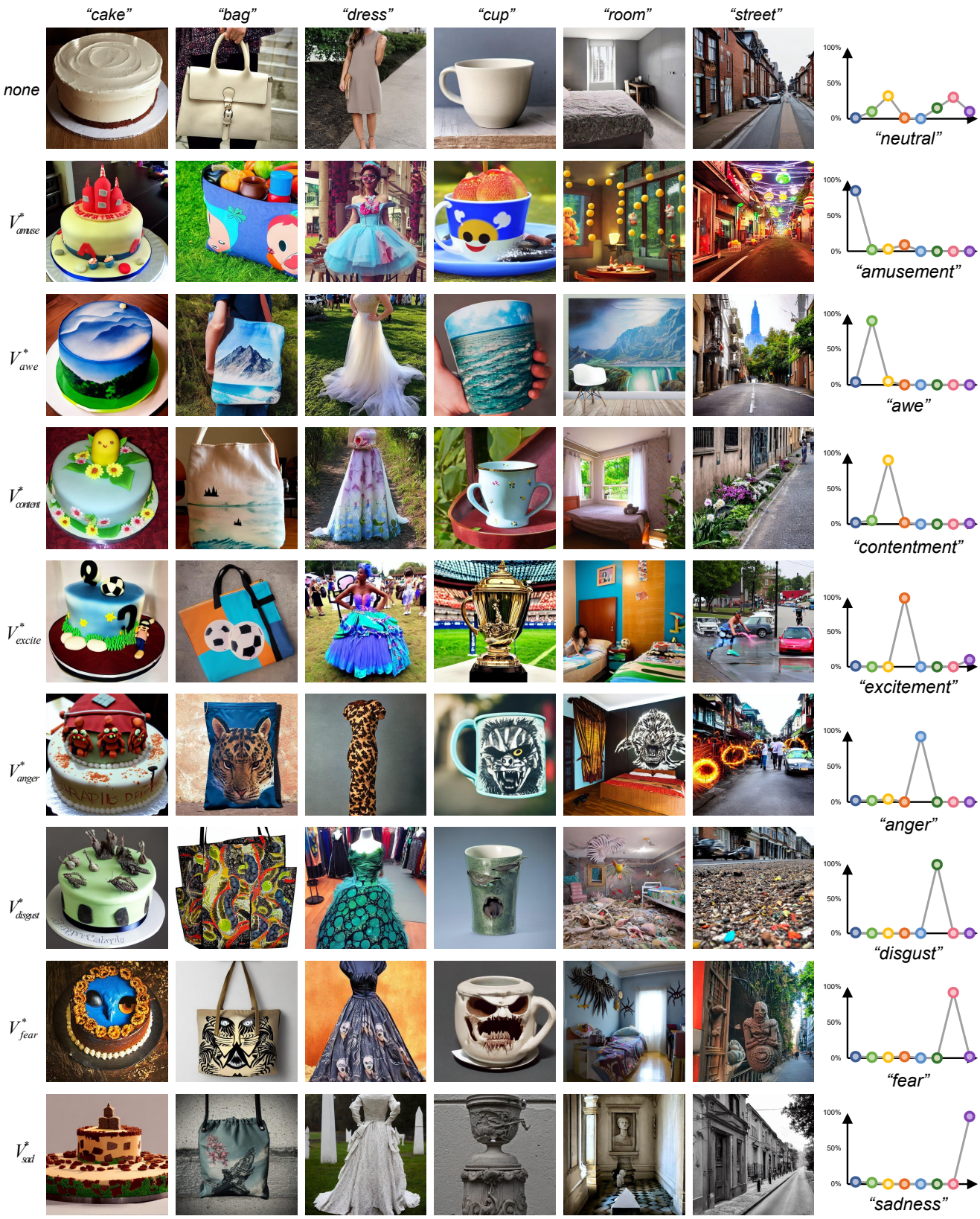

Figure 7. Emotion transfer, where we combine each of the learned emotion representation with several neutral semantics.