Few-shot Relation Personalization

Qiqian Fu
ZJU-UIUC Institute
Zhejiang University
Haining, China
qiqian.21@intl.zju.edu.cn
3210115455

Junzhou Fang
ZJU-UIUC Institute
Zhejiang University
Haining, China
junzhou.21@intl.zju.edu.cn
3210115452

Sanhe Fu
ZJU-UIUC Institute
Zhejiang University
Haining, China
sanhe.21@intl.zju.edu.cn
3210110231

Lishan Shi
ZJU-UIUC Institute
Zhejiang University
Haining, China
lishan.21@intl.zju.edu.cn
3210112476

I. Introduction

In this project, we focus on the concept of few-shot relation personalization. We are provided with several reference pictures, and these reference pictures all share a common underlying relation. The key objective is to generate new images in which the objects within them interact in accordance with this shared relation.

A. Background

We have seen great success in Text-to-Image (T2I) models recently. Given a text prompt and a few reference images, T2I models achieve a good performance in generating a required image. One of the research focuses on T2I models lies in studying the relation between entities in reference images. One of the outstanding works in this perspective is "ReVersion: Diffusion-Based Relation Inversion from Images." [1] Building on this paper, our project aims to extend the performance of the work in a few-shot scenarios.

B. Baseline

The ReVersion model consists of two components: a pre-trained frozen T2I model and a steering module to emphasize object relationships.

1) Frozen T2I model: ReVersion adapted the Stable Diffusion Model [2] where the model weights are kept frozen during the optimization process. The Stable Diffusion Model consists of a CLIP text encoder, VAE image-latent encoder, and U-Net as generative model.

The model works in two steps. During the training step, the images are iteratively corrupted with Gaussian noise during training, and the model learns to reverse this process. Then, in the inference step, the model generates an image starting from pure noise, iteratively refining it toward a clean image using learned denoising steps.

2) Steering module: The module is designed to ensure that the optimized relation prompt $< R >$ captures the high-level relational structure between objects in images, while avoiding the leakage of unrelated details such as object appearances or low-level features. First, the author used linguistic research to prove that prepositions inherently describe relations and that these prepositions can be further abstracted to a small subset to represent spatial relationships. Given this intuition, a relation-steering contrastive loss is defined to push $< R >$ toward a subspace associated with prepositions and away from irrelevant clusters. Positive samples include basis prepositions such as 'on' and 'under', and negative samples include other words that are none and verbs. The loss is further designed to tolerate the noises in the positive set by focusing on sparsely activated prepositions. This loss can be formally described by the formula:

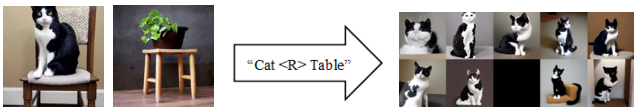

Fig. 1. One test case of using two training samples for class "on"

Final loss formula:

$$L_{steer} = -log \frac{\sum_{l=1}^{L} e^{R^\top \cdot P_i^l / \gamma}}{\sum_{l=1}^{L} e^{R^\top \cdot P_i^l / \gamma} + \sum_{m=1}^{M} e^{R^\top \cdot N_i^m / \gamma}}$$

- P_i^l : Positive prepositional samples.
- N_i^m : Negative samples.
- γ : Temperature parameter controlling sharpness.

II. Methodology(Progress)

When running the ReVersion model, we found that although the model can generate images with correct relationship between entities, it would generate strange or constant entities if the training samples are reduced. In Fig.1, we can see that if we only used two training samples, the generated images will tend to be similar and strange. The extra steering loss has moved a part of the model's attention from entities' recognition to spatial relation analysis. Therefore, when we decrease the training samples from originally ten images for one class to two images for one class, we found that although the model can still capture some spatial relationships between objects, the generated images contained strange objects or only similar objects. Fig.1 shows one test case of using only two training samples for class "on". Then we prompted the model with "Cat $< R >$ Table." One can easily observe that the output images have only one type of cat (black and white), and cats and tables are mistaken.

A. LLM Augmented Captioning

We want to weaken the model's attention for entities' detailed characteristics and let the model produce correct objects

```
Previous:
    "0.png": [
        "cat {} chair",
        "a cat {} a chair",
        "a black and white cat {} a brown chair",
        "a black and white cat {} a wooden chair, simple background"
    ],
    "1.png": [
        "pot {} stool",
        "a pot {} a stool",
        "a pot with green plant {} a wooden stool",
        "a pot with green plant {} a wooden stool, indoor"
    ],

LLM augmented:
    "0.png": [
        "cat {} chair",
        "a cat {} a chair",
        "a black and white cat {} a brown chair",
        "a black and white cat {} a wooden chair, simple background",
        "a cat sitting {} a wooden chair",
        "a cat perched {} a brown chair",
        "a black and white cat {} a chair in a neutral background",
        "a playful cat {} a chair",
        "a curious cat {} a chair"
    ],
    "1.png": [
        "pot {} stool",
        "a pot {} a stool",
        "a pot with green plant {} a wooden stool",
        "a pot with green plant {} a wooden stool, indoor",
        "a black pot {} supported by a light wooden stool",
        "a leafy plant {} inside a black pot on a stool",
        "a black pot {} resting on a wooden base"
    ]
```

Fig. 2. LLM Augmented Captioning

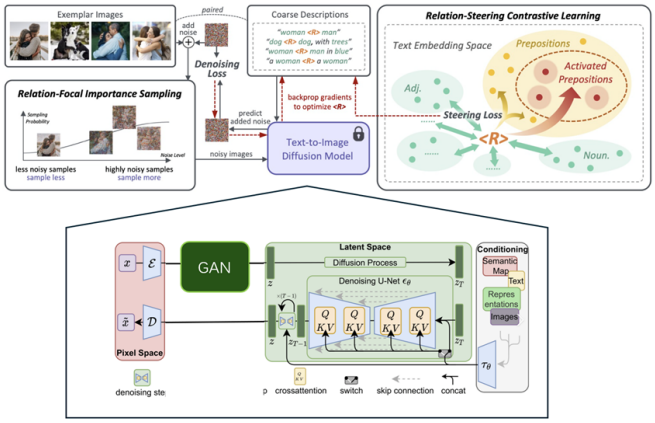

Fig. 3. Few-shot relation generation pipeline

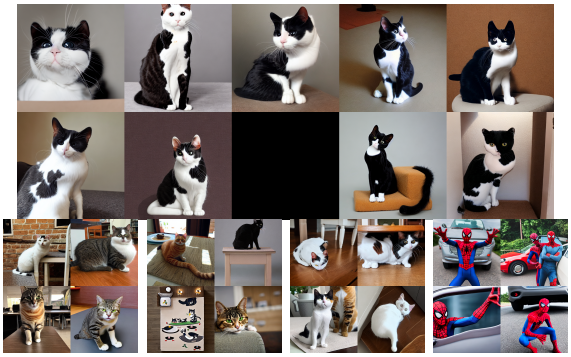

Fig. 4. Resulting images

with more variance. Therefore, we leverage LLM to generate more captions for each training image. Some examples can be seen in Fig.2.

B. GAN

We further implement a GAN module between VAE and U-Net in the frozen diffusion model. Initially, the input images will be encoded into latent space by VAE and passed to U-Net. U-Net is the central component of the diffusion process in Stable Diffusion. It operates in the latent space and performs the denoising task, progressively removing noise from the noisy latent representations to recover clean latent features. That is to say, the output of U-Net should be the predicted noise, and $\hat{z}_t = z_t - model\_pred$. UNet enables diverse outputs from the same initial latent representation, making it a flexible tool for controlled and diverse image generation. So, in this project, we treat U-Net as the generator of GAN. We add a traditional discriminator in the pipeline, during the training process, the discriminator will be trained to distinguish the latent features and predicted noise. Then, we add a GAN loss to the total loss, and the loss evaluates the ability of the discriminator to mistakenly classify the predicted noise as a real one. So the U-Net, which is the generator, will try its best to generate noise similar to latent features. By this adversarial learning scheme, we ask the diffusion model to generate more noise similar to original image, solving the difficulties of insufficient training samples. The few-shot relation generation pipeline is shown in the Fig.3.

We also ran a series of experiments to conclude that training VAE+GAN+Unet simultaneously outperforms training VAE+GAN and GAN+Unet sequentially.

III. Preliminary Results

A. Few-shot Dataset Construction

Originally, ReVersion Benchmark contains ten representative object relations, and in each relation, there are ten exemplar images. We define the few-shot case for this project as using only two training images from each relation.

B. Metric

Following the metrics in the original paper, we use the CLIP score between a revised text prompt and the generated image, which is referred to as the Entity Accuracy Score [3]. We performed the test of four versions and summarized them in the Table 1. We also listed all the resulting images in Fig.4.

The difference between Try 1 and Try 2 is the setting in the GAN network. Although the mark for single Text Augmentation is highest, we could observe from the result images that GAN + Text Aug Try 1 look best. There are several

TABLE I
THE TEST OF FOUR VERSIONS

Method	Score
Baseline:	23.96
Text Augmentation:	28.70
GAN + Text Aug Try 1:	28.25
GAN + Text Aug Try 2:	26.78

other metrics mentioned in the original paper, like Relation Accuracy Score. But these metrics were not carefully defined, so we didn't include them in this report. Given only all the result images, we can conclude that GAN + Text Aug Try 1 has the optimal performance.

REFERENCES

[1] ReVersion: Diffusion-Based Relation Inversion from Images.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[3] 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 ICML, 2021.