Supplementary Material for Textual Semantics Matters: Unsupervised Representation Disentanglement in Realistic Scenarios with Language Inductive Bias

1. More Implementation Details of Text-Aided Regularization

1.1. Text Prompts Definition

Carefully designed text prompts are crucial for the effectiveness of our text-aided regularization. In the absence of ground truth annotations of real-world datasets, we predefine all underlying factors of each dataset. For the FFHQ dataset, the factors include {hair color, hair length, bangs, head direction, head view, skin color, smile, glasses, age, gender, background color}; for the AFHQ dataset, the factors are {fur color, fur length, ear shape, nose size, mouth opening, face size, background color}; and for the Stanford Cars dataset, the factors are {model, orientation, size, car color, grille type, background color}. We design a set of structurally similar text prompts to describe the underlying factors of each dataset. The text prompts used for different datasets are provided in Table 1, Table 2 and Table 3.

1.2. Details of Attribute Manipulation

In the main paper, we frequently perform attribute manipulation on images. For example, given an image X, we shift $\mathbf{z}_{\text{sem}}$ by Δ_s along the dimension of factor f_i and reconstruct X_1. This operation is fundamental to both the second-stage training and the acquisition of traversal results. We employ DDIM inversion [3] technique to achieve attribute manipulation. Specifically, for a latent image $\mathbf{z}_0$, the attribute manipulation operation involves two steps. First we apply the inversion process:

$$\boldsymbol{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\boldsymbol{z}_t + \sqrt{\alpha_{t+1}}\left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_{sem}), \tag{1}$$

where $\mathbf{z}_{\text{sem}}$ is the disentangled representation of $\mathbf{z}_0$. After shifting a specific dimension of $\mathbf{z}_{\text{sem}}$ to obtain $\mathbf{z}^{'}_{sem}$, we perform the sampling process to obtain the attribute-edited image:

$$\boldsymbol{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\boldsymbol{z}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}^{'}_{sem}). \tag{2}$$

The inference steps for DDIM inversion is set to 20.

1.3. Algorithms of Different Losses

The text-aided regularization comprises two components, $\mathcal{L}_{push}$ and $\mathcal{L}_{pull}$, with their computational processes detailed in Algorithm 1 and Algorithm 2, respectively. For $\mathcal{L}_{push}$, the computation is performed on the target disentangled factor of current semantic direction, while for $\mathcal{L}_{pull}$, the calculation is carried out over the entire set of factors that should remain unchanged. Furthermore, we apply $\mathcal{L}_{push}$ and $\mathcal{L}_{pull}$ in two dynamic processes and introduce $\mathcal{L}_o$ and $\mathcal{L}_e$. The computation processes of Order Loss and Exchangeable Loss are illustrated in Algorithm 3 and Algorithm 4.

Algorithm 1 Compute Push Loss

Prepare: pre-trained CLIP, pre-designed text prompts, image X_1, X_2, target disentangled factor f
function LPUSH(X_1, X_2, f)
  $T \leftarrow$ GETTEXTPROMPT(f)
  $tss\_X_1 \leftarrow$ CLIPCALCULATOR(X_1, T)
  $tss\_X_2 \leftarrow$ CLIPCALCULATOR(X_2, T)
  $distance \leftarrow ||tss\_X_1, tss\_X_2||_2$
  $push\_loss \leftarrow \text{MAX}(0, margin - distance)^2$
  return $push\_loss$
end function

2. Additional Experimental Results

2.1. Subjective Results on Synthesis Datasets

We provide the qualitative results on two typical datasets: Shapes3D [1] and Cars3D [2]. As shown in Figure 1 and Figure 2, our method successfully identifies all 6 ground truth factors on Shapes3D and discovers more factors than 3 ground truth factors on Cars3D. The traversal results show that our method learns well-disentangled representations on both datasets.

Algorithm 2 Compute Pull Loss

Prepare: pre-trained CLIP, pre-designed text prompts, image X_1, X_2, factor set F
function LPULL(X_1, X_2, F)
  $pull\_loss \leftarrow 0$
  for all f_i in F **do**
    $T \leftarrow$ GETTEXTPROMPT(f_i)
    $tss\_X_1 \leftarrow$ CLIPCALCULATOR(X_1, T)
    $tss\_X_2 \leftarrow$ CLIPCALCULATOR(X_2, T)
    $distance \leftarrow ||tss\_X_1, tss\_X_2||_2$
    $pull\_loss \leftarrow pull\_loss + distance^2$
  end for
  return $pull\_loss$
end function

Algorithm 3 Compute Order Loss

Prepare: image X, factor set F, disentangled factor f_i
function COMPUTELORDER(X, F, f_i)
  $X_1 \leftarrow$ SHIFTIMAGE($X, f_i, offset$)
  $X_2 \leftarrow$ SHIFTIMAGE($X_1, f_i, -offset$)
  $L_o \leftarrow$ LPUSH(X, X_1, f_i) + LPULL($X, X_1, F - f_i$) + LPUSH(X_1, X_2, f_i) + LPULL($X_1, X_2, F - f_i$) + LPULL(X, X_2, F)
  return L_o
end function

Algorithm 4 Compute Exchangeable Loss

Prepare: image X, factor set F, disentangled factor f_i, disentangled factor f_j
function COMPUTELEXCHANGEABLE(X, F)
  $X_1 \leftarrow$ SHIFTIMAGE($X, f_i, offset$)
  $X_2 \leftarrow$ SHIFTIMAGE($X_1, f_j, offset$)
  $X_3 \leftarrow$ SHIFTIMAGE($X, f_j, offset$)
  $X_4 \leftarrow$ SHIFTIMAGE($X_3, f_i, offset$)
  $L_e \leftarrow$ LPULL(X_2, X_4, F)
  return L_e
end function

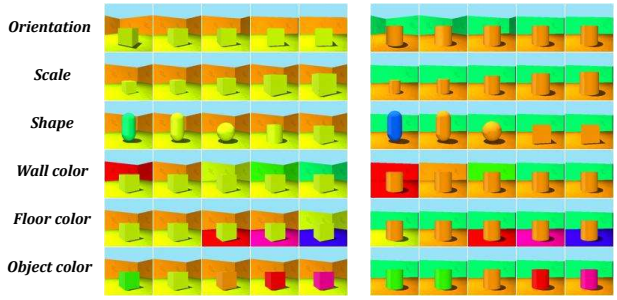

Figure 1. Latent traversal results on the Shapes3D dataset. The traversal range is (-2, 2).

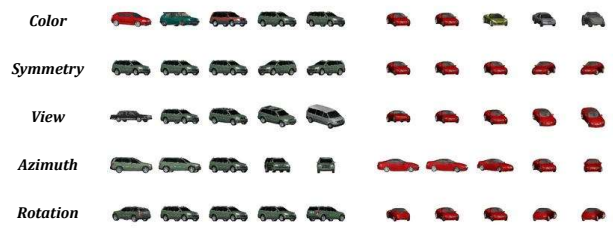

Figure 2. Latent traversal results on the Cars3D dataset. The traversal range is (-2, 2).

2.2. Image Manipulation Task Validation

To further verify the effectiveness of our TA-Dis on practical task, we conductes image manipulation experiments on different datasets. The results of manipulating on different attributes on the FFHQ, AFHQ and Stanford Cars datasets are shown in Figure 3. Benefiting from the powerful disentanglement capability of our framework, a specific attribute can be manipulated just by shifting the corresponding dimension of the latent semantic code.

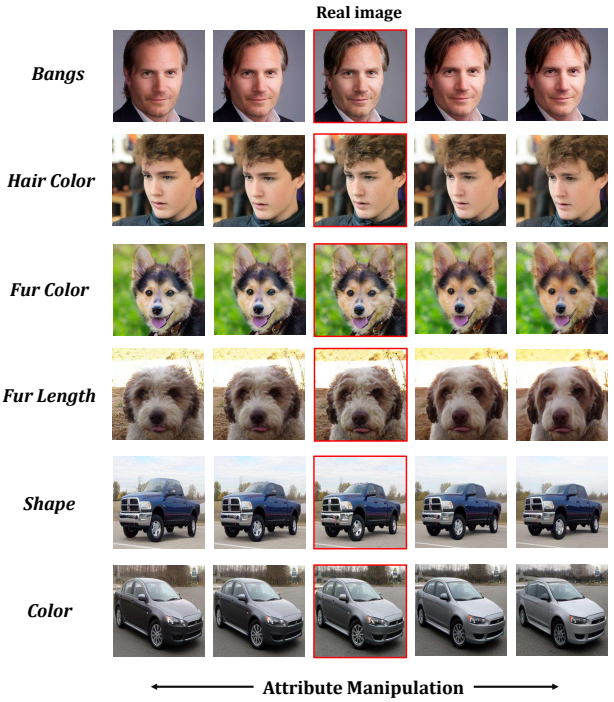

Figure 3. Practical task validation of attribute manipulation on FFHQ, AFHQ and Stanford Car datasets. With the aid of text, these attributes align well with their corresponding dimensions. Only shifting a single dimension results in effective attribute manipulation performance.

3. Limitation Discussions

As described in Section **??**, we predefine the correspondence between a specific dimension of the latent semantic code and an underlying factor of real-world scenarios to effectively apply text-aided regularization, rather than relying automatic semantic alignment. This approach is adopted because the dimension-wise preliminary disentanglement capabilities of the first-stage pre-trained model provide an essential basis for applying semantic alignment, despite its suboptimal generation ability. Moreover, abandoning the first-stage pre-training is infeasible, as high-quality images are necessary for CLIP. We will explore potential ways to overcome this limitation in future work.

factor	prompts
model	The model of the car is a sedan.
	The model of the car is a van.
	The model of the car is an SUV.
	The model of the car is a sports car.
	The model of the car is a pickup truck.
	The model of the car is a hatchback.
	The model of the car is a coupe.
orientation	The orientation of the car is facing forward.
	The orientation of the car is angled to the left.
	The orientation of the car is angled to the right.
	The orientation of the car is from the side.
	The orientation of the car is facing away.
size	The size of the car is compact.
	The size of the car is midsize.
	The size of the car is full-size.
	The size of the car is large, like an SUV.
	The size of the car is extra-large, like a truck.
car color	The color of the car is black.
	The color of the car is white.
	The color of the car is red.
	The color of the car is blue.
	The color of the car is silver.
	The color of the car is green.
	The color of the car is yellow.
	The color of the car is gray.
grille type	The grille of the car is horizontal.
	The grille of the car is vertical.
	The grille of the car is honeycomb-style.
	The grille of the car is mesh.
	The grille of the car is small to none.
background color	The background color of the image is blue.
	The background color of the image is green.
	The background color of the image is red.
	The background color of the image is yellow.
	The background color of the image is white.

Table 1. Text prompts on the Stanford Cars dataset.

factor	prompts
hair color	A person with black hair.
	A person with brown hair.
	A person with blonde hair.
	A person with red hair.
	A person with gray hair.
	A person with blue hair.
hair length	A person with a shaved head, having no visible hair.
	A person with very short hair, close to the scalp.
	A person with medium-length hair reaching just above the shoulders.
	A person with long hair flowing down to the shoulders.
	A person with very long hair extending below the waist.
bangs	A person without bangs, having a clean forehead visible.
	A person with short bangs, covering half of the face.
	A person with medium length bangs, reaching just above the eyes.
	A person with long bangs, falling past the eyes.
head direction	A person with head turned to the left.
	A person with head turned to the right.
	A person with head looking straight ahead.
head view	A person with their head fully tilted upwards.
	A person with their head slightly tilted upwards, gazing above eye level.
	A person with their head held straight, facing directly forward.
	A person with their head slightly tilted downwards, looking towards the ground.
skin color	A person with yellow skin color.
	A person with black skin color.
	A person with white skin color.
	A person with brown skin color.
smile	A person with a serious expression, not smiling.
	A person slightly smiling without showing teeth.
	A person showing a big smile with teeth exposed.
	A person laughing joyfully with mouth open and teeth visible.
glasses	A person with no eyeglasses.
	A person wearing metal-framed eyeglasses with transparent lenses.
	A person wearing sunglasses with light-colored lenses.
	A person wearing sunglasses with dark lenses.
age	A child under 15 years old.
	A teenager or young adult, aged between 15 and 30.
	A middle-aged adult, between 30 and 60 years old.
	An older adult, aged above 60.
gender	A person presenting as male.
	A person presenting as female.
	A person with an androgynous appearance, presenting traits of both male and female.
background color	The background color of the image is blue.
	The background color of the image is green.
	The background color of the image is red.
	The background color of the image is yellow.
	The background color of the image is white.

Table 2. Text prompts on the FFHQ dataset.

factor	prompts
fur color	The fur color of the dog is white.
	The fur color of the dog is black.
	The fur color of the dog is brown.
	The fur color of the dog is golden.
	The fur color of the dog is gray.
	The fur color of the dog is a mix of black and white patches.
	The fur color of the dog is a blend of brown and white.
fur length	The fur length of the dog is very short.
	The fur length of the dog is short.
	The fur length of the dog is medium.
	The fur length of the dog is long.
	The fur length of the dog is very long.
ear shape	The ear shape of the dog is floppy.
	The ear shape of the dog is pointy.
	The ear shape of the dog is semi-floppy.
	The ear shape of the dog is short and rounded.
	The ear shape of the dog is large and upright.
nose size	The nose size of the dog is small.
	The nose size of the dog is medium.
	The nose size of the dog is large.
	The nose size of the dog is broad.
	The nose size of the dog is narrow.
mouth opening	The mouth opening of the dog is closed.
	The mouth opening of the dog is slightly open.
	The mouth opening of the dog is half open.
	The mouth opening of the dog is wide open.
	The mouth opening of the dog is fully open with tongue out.
face size	The face size of the dog is small.
	The face size of the dog is medium.
	The face size of the dog is large.
	The face size of the dog is very large.
background color	The background color of the image is blue.
	The background color of the image is green.
	The background color of the image is red.
	The background color of the image is yellow.
	The background color of the image is white.

Table 3. Text prompts on the AFHQ dataset.

References

[1] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 1

[2] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015. 1

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1