ICCV #12241

Supplementary Material for "General Image-to-Image Translation with One-Shot Image Guidance"

Anonymous ICCV submission

Paper ID 12241

1. More Implentation Details

Algorithm 1 Visual Concept Translator

Require: x^{src}, x^{ref}: source and reference image.
Require: α, β: learning rates; S_m, S_p: training steps.

Initial content embedding v^{src} and concept embedding v^{ref}.
▷ Multi-concept inversion
for step=1,...,S_m **do**
  Compute L_{ldm} and L_{rec} in Eq. (14) and Eq. (15);
  Update concept embedding with gradient descent: $v^{ref} \longleftarrow v^{ref} - \alpha \bigtriangledown_{v^{ref}} (L_{ldm} + L_{rec})$.
end for
▷ Pivotal turning inversion
Compute source latent $z^{src} = \mathcal{E}(x^{src})$ with encoder $\mathcal{E}$;
Compute z_T =DDIM-Inversion(z^{src}) with Eq. (4)
Compute unconditional embedding $v_\varnothing = \tau(\varnothing)$ with tokenizer τ;
for t=T,...,1 **do**;
  for step=1,...,S_p **do**
    Compute $L = z_0 - \hat{z}_0 \left(z_t, v_t^{src}\right)$ in Eq. (12);
    Update content embedding with gradient descent: $v_t^{src} \longleftarrow v_t^{src} - \alpha \bigtriangledown_{v_t^{src}} L$
  end for
  $\hat{\epsilon} \leftarrow \tilde{\varepsilon}_\theta \left(z_t, t, v_t^{src}, v^\varnothing\right)$ in Eq. (10);
  $z_{t-1} \leftarrow$ DDIM-sample$(z_t, \hat{\epsilon}, t)$
end for
▷ Content-concept fusion
for t=T,...,1 **do**;
  Compute noise prediction ϵ^{src} and ϵ^{ref} in Eq. (7);
  $M_t, \hat{\epsilon} \leftarrow \tilde{\varepsilon}_\theta \left(z_t, t, v_t^{src}, v^{ref}\right)$ in Eq. (8);
  $M_t^*, \hat{\epsilon}^* \leftarrow \tilde{\varepsilon}_\theta \left(z_t^*, t, v_t^{src}, v^\varnothing\right)$ in Eq. (10);
  $\widehat{M}_t \leftarrow AC\left(M_t, M_t^*, t\right)$ in Eq. (11)
  $\hat{\epsilon} = \tilde{\varepsilon}_\theta \left(z_t, t, v_t^{src}, v^{ref}\right) \{M \leftarrow \widehat{M}_t\}$
  $z_{t-1} \leftarrow$ DDIM-sample$(z_t, \hat{\epsilon}, t)$
  $z_{t-1}^* \leftarrow$ DDIM-sample$(z_t^*, \hat{\epsilon}^*, t)$
end for
Compute target image $x^{tgt} = \mathcal{D}(z^0)$ with decoder $\mathcal{D}$;

Our full algorithm is shown in Algorithm 1. For multi-concept inversion, we empirically found that 200 training steps are enough for convergence, and this process only takes about 150 seconds. Furthermore, for pivotal turning inversion, our found optimal training step is 1000, which takes about 60 seconds. The learning rate is 5×10^{-4} for multi-concept inversion. For pivotal turning inversion, we reduce the learning rate when step increases, as

$$lr = 1 \times 10^{-2} \times s / 5000, \tag{1}$$

where s is the current step numbers. The algorithm also includes the unconditional embedding $v_\varnothing$, which is extracted by putting empty text to the BERT tokenizer. The Adam optimizer is used for both inversion processes.

2. More Results of General Image-to-image Translation

To further verify the model performance in the general image-to-image translation tasks, we make more experiments with different reference images, as shown in Fig. 1.

It's noted that there is a trade-off between structural preservation and semantic changes. As shown in Fig. 2, the injection ratio of cross-attention and self-attention affects the result a lot. To obtain the ideal results, we can adjust the attention injection ratio to the optimal value. Empirically, we adopt a low cross-attention injection ratio of about 20%, and adjust the self-attention injection ratio to achieve different preservation results.

3. More Results of Style Transfer

To further verify the model performance in the style transfer task, we make more experiments with different reference styles, as shown in Fig. 3. In the figure, the first column contains the reference images, and the first row contains the content images. The model outputs show the excellent performance of the proposed method with content preserved and style transferred.

As a type of style transfer, portrait style transfer tries to substitute the input face with another stylized face. The

ICCV #12241

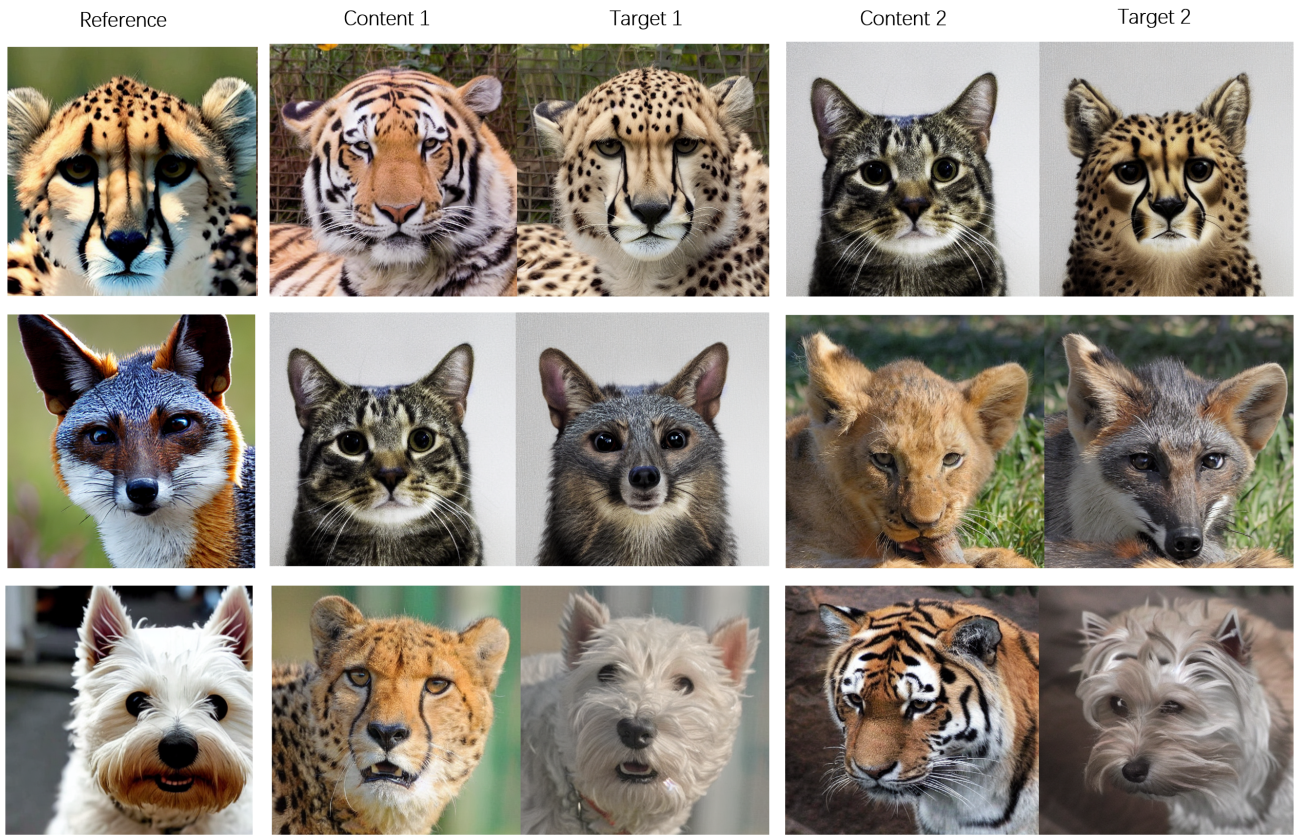

Figure 1: Model performance in general image-to-image translation tasks. The first column contains the reference images, and the following columns contain two groups of results based on the content images in column 2 and column 4. Our model generates realistic samples that reflect the reference image while maintaining the structure of the source image.

proposed algorithm shows excellent performance in the task of portrait style as Fig. 4. Given the one-shot input, our model can substitute the face in the reference style with the face in the content image with high quality.

4. More Ablation Study

We evaluate the influence of the number of multi-concept embeddings, as described in part of multi-concept inversion (Section 3.4) in the main paper. Given a reference image, we visualize the model performance with different concept embeddings as shown in Fig. 5. From the figure, a small embedding number cannot well translate the concepts in the reference image, as in columns 2-3. A too-large embedding number still leads to poor performance with translation failures, as in columns 5-6. We empirically found that using 3 concept embeddings is the best choice, as in column 4.

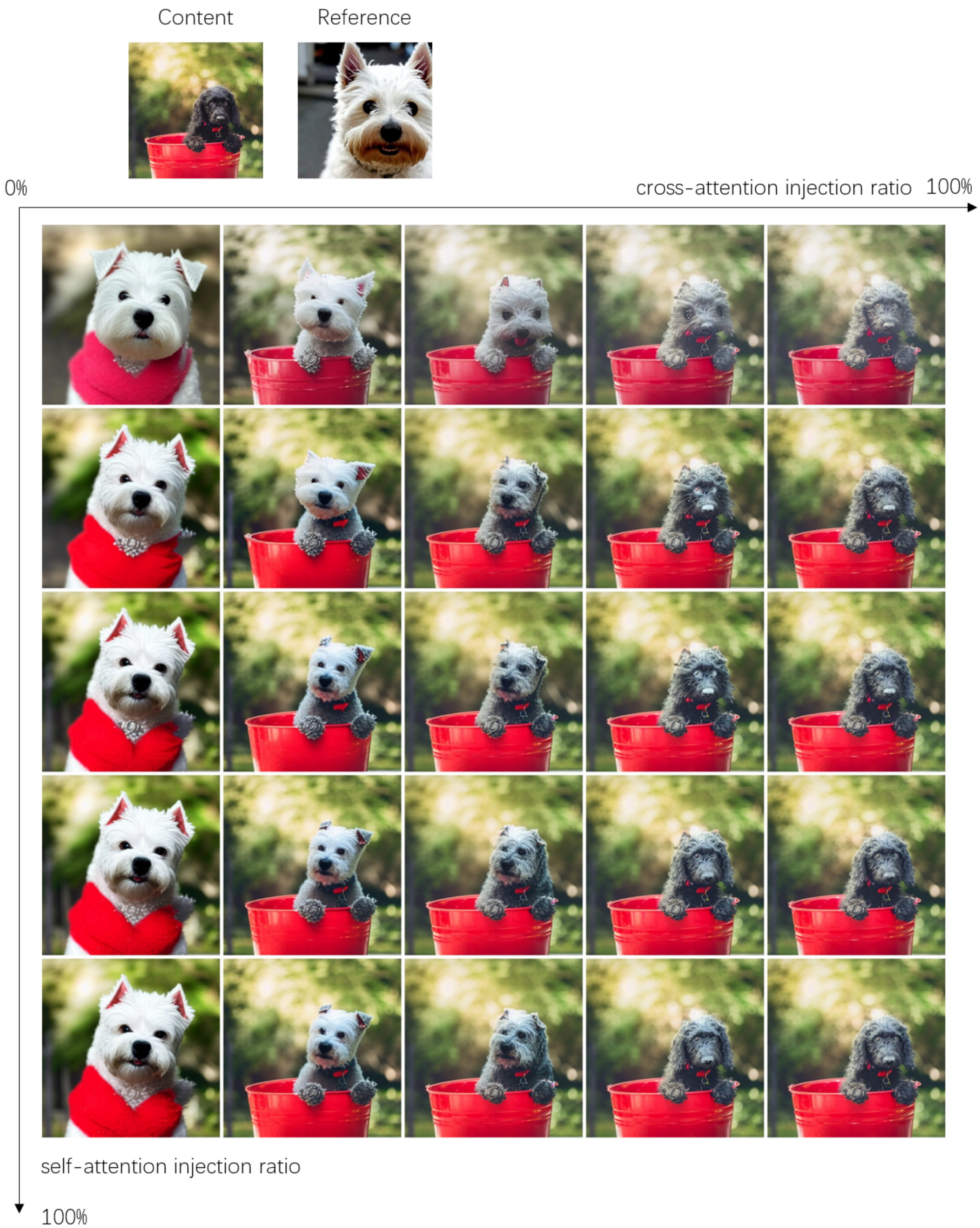

Figure 2: Trade-off between structural preservation and semantic changes. We generate the ideal results by adjusting the cross-attention and self-attention injection ratios to optimal values.

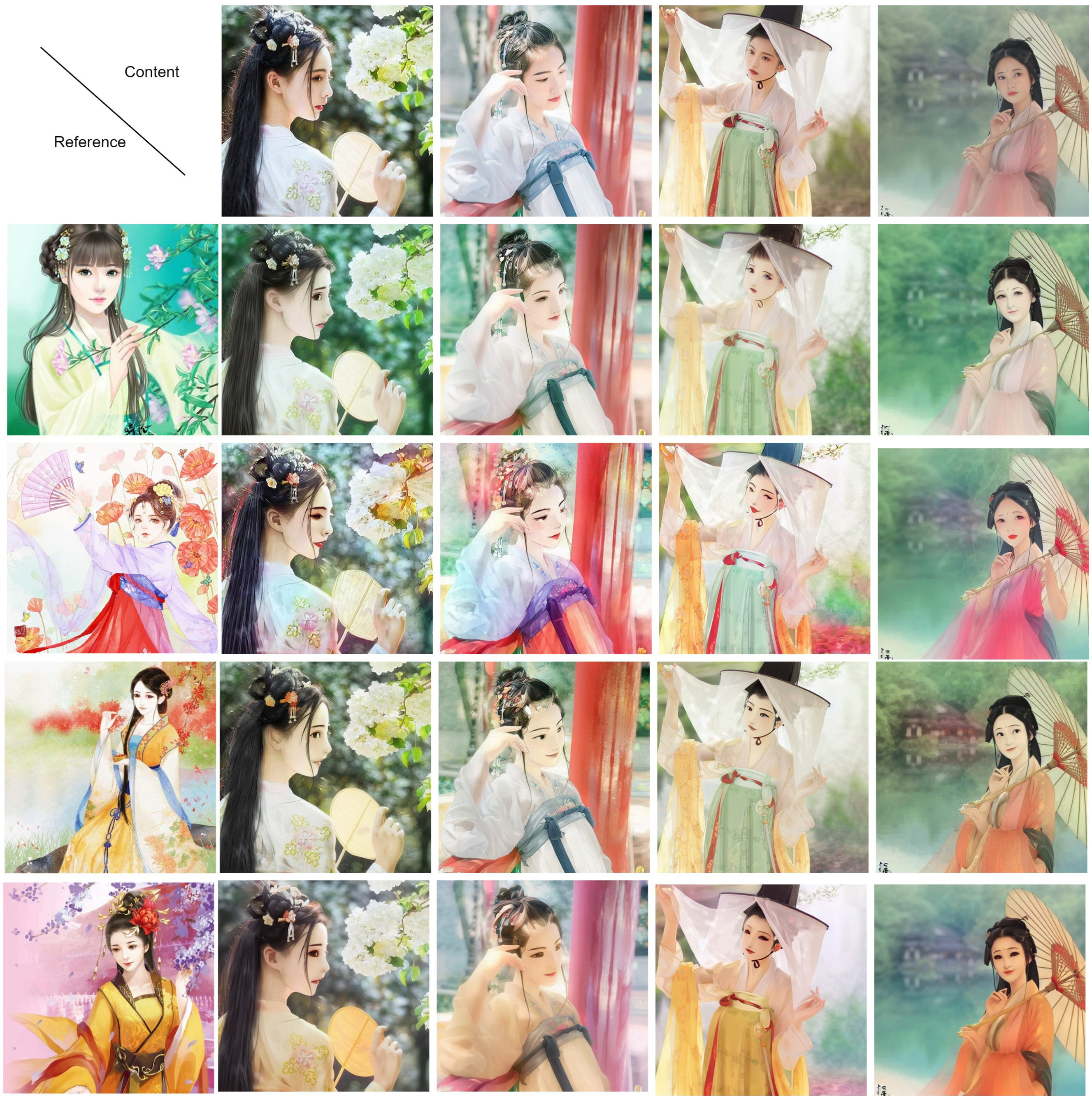

Figure 3: Model performance in style transfer tasks. The first column contains the reference images, and the first row contains the content images. The other images are the model outputs based on corresponding content and reference images.

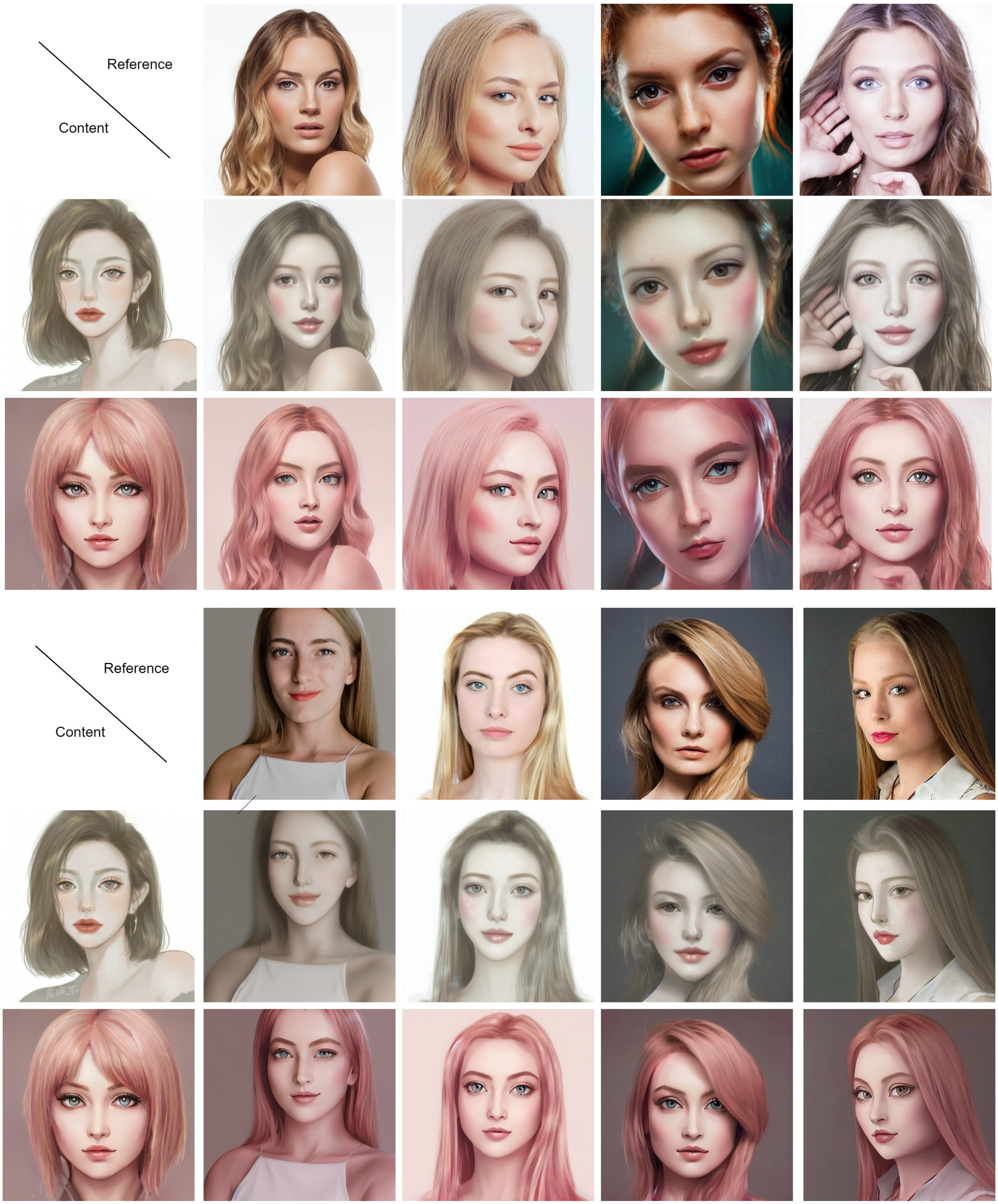

Figure 4: The model performance on portrait style transfer. Given the one-shot input, our model can substitute the face in the reference style image with the face in the content image with high quality.

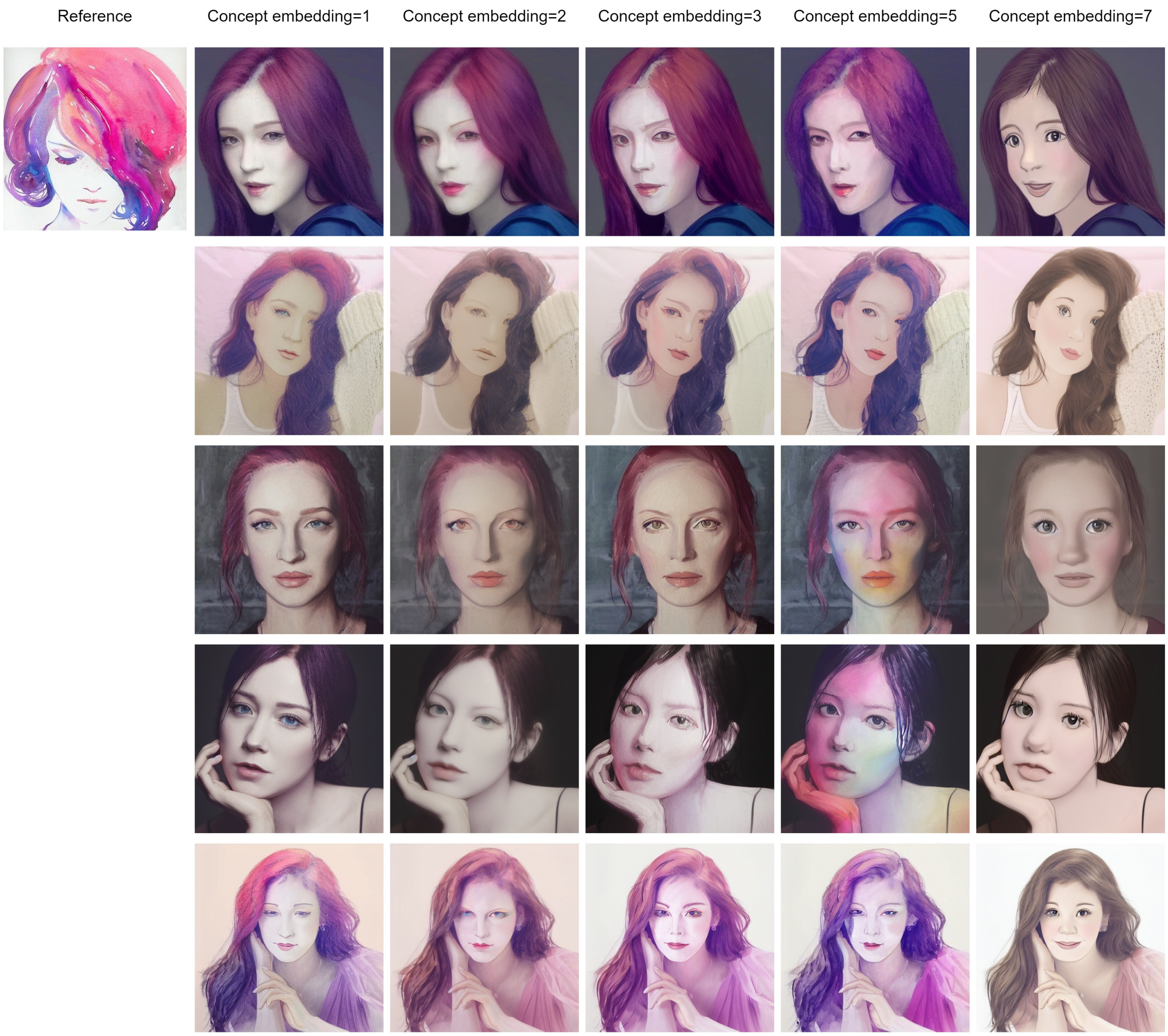

Figure 5: The model performance with different numbers of concept embedding. A small embedding number cannot well translate the concepts in the reference image, as in columns 2-3. A too-large embedding number still leads to poor performance with translation failures, as in columns 5-6. We empirically found that using 3 concept embeddings is the best choice, as in column 3.