Supplementary Materials:
ToonTalker: Cross-Domain Face Reenactment

A. Network Architecture Details

The architectures of several key model modules are shown in Fig. 1. Please note that the module structures of the real and cartoon domains are the same. "ResBlock-64" denotes the number of feature map channels in the residual block is 64.

Appearance encoder E_a. As shown in Fig. 1(a), our appearance encoder consists of six ResBlocks. In each block the size of feature maps will be downsampled. Thus we can obtain feature maps with the different sizes, *i.e.*, 8×8, 16×16, 32×32, 64×64, 128×128, and 256×256.

Motion encoder E_m. As shown in Fig. 1(b), the motion encoder is composed of seven ResBlocks. And we use a 4×4 convolution to downsample the feature map after the sever ResBlocks. The output motion code can be denoted as $F \in R^{512}$.

Motion Query Transformer T. We use transformers to align motion spaces with common query tokens for their ability of catching long-range dependencies. In our model, as shown *in Fig.3 of the manuscript*, there are two query transformers, *i.e.,* source query transformer and driving query transformer. The architectures of the two transformers are the same. The motion base consists of 20 learnable embeddings, denoted as $p = \{p^k\}_{k=1}^{K}, K = 20$ and $p^k \in R^d, d = 512$.

Backward Transformer T_B. As shown *in Fig.3 of the manuscript*, the backward transformer consists of three transformer blocks, which correspond to multi-scale feature maps.

Generator G. StyleConv blocks are proved to be effective in StyleGAN. So similar to the architecture of StyleGAN, the generator contains six StyleConv blocks. In the first three blocks, the low-resolution feature map from E_a will be warped and upsampled on the condition of the motion code. And in the last three blocks, we not only warp the feature map but also refine the warped feature by spatial feature transformation. The two types of modified StyleConv blocks are shown in Fig. 1(c) and Fig. 1(d).

B. More Qualitative Results.

We show more examples of comparison with *state-of-the-art* methods on cross-domain reenactment in Fig 2.

C. Evaluation on Images Generated by Stable Diffusion

The fantastic generation ability of diffusion models has set off an upsurge in the world. Therefore, we use several cartoon characters generated by diffusion models as source images for cross-domain reenactment. We show some samples animated by our method in Fig. 3.

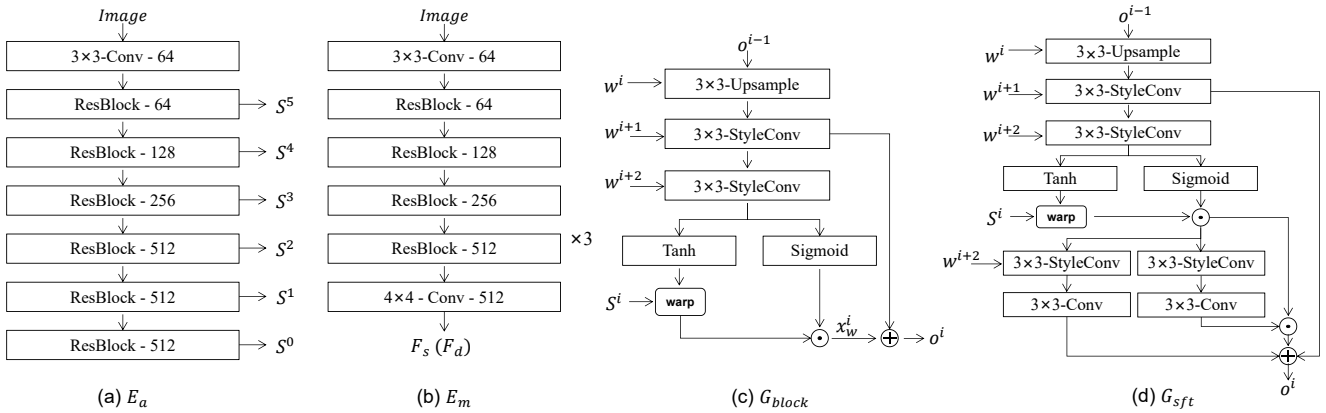

Figure 1. The architecture of our ToonTalker.

Figure 2. Qualitative comparisons with *state-of-the-art* methods on cross-domain reenactment.

Figure 3. Animating characters generated by Stable Diffusion.