# Supplementary Materials

Bolin Chen, Zhao Wang, Binzhe Li, Shurun Wang, Shiqi Wang, *Senior Member, IEEE* and Yan Ye, *Senior Member, IEEE*

**Abstract**—This is the supplementary material for the paper entitled "Interactive Face Video Coding: A Generative Compression Framework". First, we provide an overall introduction regarding model training. Besides, we supplement the detailed results of rate-distortion performance, subjective performance, ablation study and interactive coding examples. The project page can be found at https://github.com/Berlin0610/Interactive_Face_Video_Coding.

**Index Terms**—Interactive video coding, controllable embedding, face video.

## 1 Model Training

During our model training, the self-supervised training strategy is adopted to jointly optimize the mesh-based motion estimation and frame generation modules. The loss objectives in our model training mainly include perceptual loss, adversarial loss, identity loss, flow loss, texture loss and pixel loss. Herein, we provide the details of all loss functions.

### 1.1 Perceptual Loss

We employ the perceptual loss [1] based on the pre-trained VGG-19 network to improve the visual quality of reconstructed talking face images. This loss term is used twice in our model training, where the reconstruction of coarse deformed frame $\mathcal{F}_{cdf}^{I_l}$ and final prediction result $\hat{I}_l$ are supervised with the help of original inter frame $I_l$, respectively. Let $VGG_i \in R^{C_i \times H_i \times W_i}$ be the feature map of the $i_{th}$ layer of VGG-19 network, these two perceptual loss functions (*i.e.*, $\mathcal{L}_{per1}$ and $\mathcal{L}_{per2}$) can be described as follows,

$$\mathcal{L}_{per1} = \sum_{i=1}^{5} \frac{\left\| VGG_i\left(\mathcal{F}_{cdf}^{I_l}\right) - VGG_i(I_l) \right\|}{C_i \times H_i \times W_i}, \tag{1}$$

$$\mathcal{L}_{per2} = \sum_{i=1}^{5} \frac{\left\| VGG_i\left(\hat{I}_l\right) - VGG_i(I_l) \right\|}{C_i \times H_i \times W_i}. \tag{2}$$

### 1.2 Adversarial Loss

We use the patch GAN implemented in [2], [3] and the hinge loss to generate realistic image manifold for talking face reconstruction. In analogous to the perceptual loss training, this adversarial loss term is also used twice for coarse deformed frame $\mathcal{F}_{cdf}^{I_l}$ and final prediction frame $\hat{I}_l$. The discriminator $D$, including multi-scale discriminators (*i.e.*, $D_1, D_2, D_3$ and $D_4$), are operated on different frame resolutions. Besides, the feature matching loss [2] is jointly used for stable model training. The loss functions of the generator $G$ and discriminator $D$ are represented as follows,

$$\mathcal{L}_{G1}\left(\mathcal{F}_{cdf}^{I_l}\right) = -\sum_{k=1}^{4} E_{\mathcal{F}_{cdf}^{I_l} \sim P_g}\left[D_k\left(\mathcal{F}_{cdf}^{I_l}\right)\right], \tag{3}$$

$$\begin{aligned}\mathcal{L}_{D1}\left(\mathcal{F}_{cdf}^{I_l}, I_l\right) = \sum_{k=1}^{4} (&-E_{I_l \sim P_r}\left[\min\left(D_k\left(I_l\right) - 1, 0\right)\right] \\ &-E_{\mathcal{F}_{cdf}^{I_l} \sim P_g}\left[\min\left(-D_k\left(\mathcal{F}_{cdf}^{I_l}\right) - 1, 0\right)\right]),\end{aligned} \tag{4}$$

$$\mathcal{L}_{G2}\left(\hat{I}_l\right) = -\sum_{k=1}^{4} E_{\hat{I}_l \sim P_g}\left[D_k\left(\hat{I}_l\right)\right], \tag{5}$$

$$\begin{aligned}\mathcal{L}_{D2}\left(\hat{I}_l, I_l\right) = \sum_{k=1}^{4} (&-E_{I_l \sim P_r}\left[\min\left(D_k\left(I_l\right) - 1, 0\right)\right] \\ &-E_{\hat{I}_l \sim P_g}\left[\min\left(-D_k\left(\hat{I}_l\right) - 1, 0\right)\right]),\end{aligned} \tag{6}$$

where $P_g$ and $P_r$ represent the generated and real image distributions. As such, the final adversarial loss is given by,

$$\mathcal{L}_{adv1} = \mathcal{L}_{G1}\left(\mathcal{F}_{cdf}^{I_l}\right) + \mathcal{L}_{D1}\left(\mathcal{F}_{cdf}^{I_l}, I_l\right), \tag{7}$$

$$\mathcal{L}_{adv2} = \mathcal{L}_{G2}\left(\hat{I}_l\right) + \mathcal{L}_{D2}\left(\hat{I}_l, I_l\right). \tag{8}$$

### 1.3 Flow Loss

We employ the flow loss to minimize the characterization error of dense motion field between the ground-truth flow $\Gamma_{original}^{I_l}$ and the estimated dense flow $\Gamma_{fine}^{I_l}$, further improving the warping correctness during the generation process. It should be mentioned that the end-to-end spatial pyramid network (SpyNet) [4] is used to predict the ground-truth flow $\Gamma_{original}^{I_l}$. This loss is given by,

$$\mathcal{L}_{flow} = \left\| \Gamma_{original}^{I_l} - \Gamma_{fine}^{I_l} \right\|. \tag{9}$$

### 1.4 Pixel Loss

We compute the $L1$ distance as the pixel loss between final prediction frame $\hat{I}_l$ and original inter frame $I_l$, where the distance is computed as the sum of differences of every image pixel. It can be described as follows,

$$\mathcal{L}_{pixel} = \left\| I_l - \hat{I}_l \right\|. \tag{10}$$

### 1.5 Texture Loss

To better capture the texture information of generated images, the Gram matrix [5] $\mathbf{Gram}$ is introduced to calculate the feature correlations between final prediction frame $\hat{I}_l$ and original inter frame $I_l$, where the feature extraction network is also VGG-19. The corresponding loss can be formulated by,

$$\mathcal{L}_{tex} = \sum_{i=1}^{5} \frac{\left\| \mathbf{Gram}\left(VGG_i\left(\hat{I}_l\right)\right) - \mathbf{Gram}\left(VGG_i(I_l)\right) \right\|}{C_i \times H_i \times W_i}. \quad (11)$$

### 1.6 Identity Loss

We adopt the identity loss [6] to enforce the generated face with high fidelity and strong identity perseverance. In analogous to perceptual loss, face recognition model ArcFaceNet [7] is introduced to replace the VGG-19 network, such that the face identity difference between final prediction frame $\hat{I}_l$ and original inter frame $I_l$ can be minimized. Let $Arc_i \in R^{C_i \times H_i \times W_i}$ be the feature map of the $i_{th}$ layer of ArcFaceNet network. This identity loss is represented as follows,

$$\mathcal{L}_{id} = \sum_{i=1}^{4} \frac{\left\| Arc_i\left(\hat{I}_l\right) - Arc_i(I_l) \right\|}{C_i \times H_i \times W_i}. \quad (12)$$

To conclude, the overall training loss can be summarized as follows,

$$\begin{aligned} \mathcal{L}_{total} &= \lambda_{per1}\mathcal{L}_{per1} + \lambda_{per2}\mathcal{L}_{per2} + \lambda_{adv1}\mathcal{L}_{adv1} \\ &+ \lambda_{adv2}\mathcal{L}_{adv2} + \lambda_{flow}\mathcal{L}_{flow} + \lambda_{pixel}\mathcal{L}_{pixel} \\ &+ \lambda_{tex}\mathcal{L}_{tex} + \lambda_{id}\mathcal{L}_{id}, \end{aligned} \quad (13)$$

where the loss hyper-parameters are set as follows: $\lambda_{per1} = 10$, $\lambda_{adv1} = 1$, $\lambda_{per2} = 10$, $\lambda_{adv2} = 1$, $\lambda_{flow} = 20$, $\lambda_{pixel} = 100$, $\lambda_{tex} = 100$ and $\lambda_{id} = 40$.

## 2 Experimental Results

### 2.1 Comparison Methods

To verify the performance of our proposed face interactive coding scheme, we employ the latest hybrid video coding standard VVC [8] and five generative compression schemes, including FOMM [9], FOMM2.0 [10], CFTE [11], Face_vid2vid [12] and Face2FaceRHO [13] as anchors. In the following, we discuss the implementation details of these anchors.

#### *2.1.1 Traditional VVC Codec*

It is the latest hybrid video coding standard, which significantly improves the rate-distortion performance compared with its predecessors. We adopt the Low-Delay-Bidirectional (LDB) configuration in VTM 10.0 reference software for VVC, where the quantization parameters (QP) are set to 45, 47, 50 and 52.

#### *2.1.2 Generative compression algorithms*

They are rooted in GAN-based image animation models. In our experiment, FOMM [9], FOMM2.0 [10], Face2FaceRHO [13], Face_vid2vid [12] and CFTE [11] are used as anchors. For these generative compression schemes, they strictly follow the encoder-decoder architecture. At the encoder side, the compact feature representations (e.g., 2D keypoints, 3D keypoints and 3DMM parameters) extracted from talking face frames are inter-predicted, quantized and entropy-encoded into the bitstream. When the bitstream is received at the decoder side, these compact features can be further decoded to animate the reconstructed key-reference frame and generate high-quality talking face frames. The detailed implementations are described as follows,

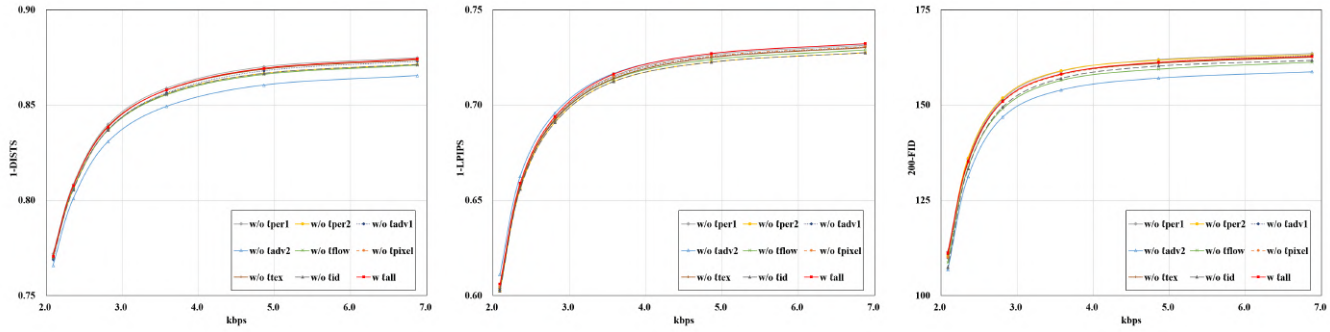
- FOMM adopts 10 groups of learned 2D keypoints $\mathbb{R}^{2\times10}$ along with their local affine transformations $\mathbb{R}^{2\times2\times10}$ to characterize complex motions. The total number of encoding parameters for each talking face frame is 60.
- FOMM2.0 extracts consistent regions of talking face to describe locations, shape, and pose, mainly represented with shift matrix $\mathbb{R}^{2\times10}$, covar matrix $\mathbb{R}^{2\times2\times10}$ and affine matrix $\mathbb{R}^{2\times2\times10}$. As such, the total number of encoding parameters for each talking face frame is 100.
- Face2FaceRHO is a 3DMM-assisted warping-based face reenactment algorithm, where 3DMM can disentangle talking face into a series of semantic parameters, including expression parameters $\mathbb{R}^{50}$, head rotation parameters $\mathbb{R}^{6}$ and head translation parameters $\mathbb{R}^{3}$. As such, the total number of encoding parameters for each talking face frame is 59.
- Face_vid2vid can estimate 12-dimension head parameters (i.e., rotation matrix $\mathbb{R}^{3\times3}$ and translation parameters $\mathbb{R}^{3}$ ) and 15 groups of learned 3D keypoint perturbations $\mathbb{R}^{3\times15}$ due to facial expressions, where the total number of encoding parameters for each talking face frame is 57.
- CFTE can model the temporal evolution of faces into learned compact feature representation with the matrix $\mathbb{R}^{4\times4}$, where the total number of encoding parameters for each talking face frame is 16.

### 2.2 RD Performance

Table 1 shows the detailed bit-rate savings of each talking face sequence. Compared with different compression algorithms, our proposed compression scheme is able to achieve advantageous bit-rate savings. In particular, our proposed scheme can achieve 75.37% average bit-rate savings in terms of DISTS, 70.29% bit-rate savings in terms of LPIPS and 74.02% bit-rate savings in terms of FID in comparison with the latest VVC codec.

### 2.3 Subjective Performance

Fig. 2 provides more visual examples of different face compression algorithms at similar bit rates. It can be noticed that our proposed interactive face coding scheme owns obvious advantages in capturing global and local facial motion, and high-quality face image reconstruction at ultra-low bit rate compared with other face compression algorithms.




(a) Rate-DISTS (b) Rate-LPIPS (c) Rate-FID

Fig. 1: Rate-distortion performance contributions of training loss functions in terms of Rate-DISTS, Rate-LPIPS and Rate-FID.

### 2.4 Ablation Study

In order to quantify the effect of each loss function during the model training, the ablation study is further executed in terms of these loss functions (e.g. $\mathcal{L}_{per1}$, $\mathcal{L}_{per2}$, $\mathcal{L}_{adv1}$, $\mathcal{L}_{adv2}$, $\mathcal{L}_{flow}$, $\mathcal{L}_{pixel}$, $\mathcal{L}_{tex}$ and $\mathcal{L}_{id}$). Fig. 1 illustrates that when all loss functions are employed to jointly optimize our proposed model, the rate-distortion performance can be all improved in terms of Rate-DISTS, Rate-LPIPS and Rate-FID.

### 2.5 Interactive Face Video Coding

Our proposed compression framework enjoys the best aspects of compact representation and semantic interpretation by characterizing highly-independent facial semantics from face frames. As such, it is of great benefits to actualize controllable semantic interactivity or virtual character animation based on these facial semantics.

#### *2.5.1 Controllable manipulation for friendly interactivity*

Fig. 3 and Fig. 4 provide more visual examples regarding face interactive coding. As shown in Fig. 3, our proposed face interactive coding scheme has great flexibility in controlling eye blinking, mouth motion, head rotation and head translation. As shown in Fig. 4, our proposed scheme has strong degree of freedom in interactivity such that the eye motion, mouth motion, head rotation and head translation of face videos can be randomly controlled. Such a compression mechanism can well adapt to future video conferencing in terms of interactivity.

#### *2.5.2 Virtual character animation for privacy protection*

Fig. 5 provides more visual examples of virtual character animation based on compact facial semantics. Experimental results have demonstrated the robustness of our proposed algorithm since it can animate any cross-identity face character and reconstruct face video with accurate pose and expression. As such, this mechanism provides great promises for virtual live entertainment and privacy-protected video conferencing.

## REFERENCES

[1] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[3] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[4] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2720–2729.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[6] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4685–4694.

[8] B. Bross, Y. Wang, Y. Ye, S. Liu, J. Chen, G. Sullivan, and J. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[9] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, pp. 7137–7147, 2019.

[10] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 648–13 657.

[11] B. Chen, Z. Wang, B. Li, R. Lin, S. Wang, and Y. Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *Proceedings of the IEEE Data Compression Conference*, 2022, pp. 13–22.

[12] T. Wang, A. Mallya, and M. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 039–10 049.

[13] K. Yang, K. Chen, D. Guo, S.-H. Zhang, Y.-C. Guo, and W. Zhang, "Face2face:real-time high-resolution one-shot face reenactment," in *European Conference of Computer vision*, 2022.

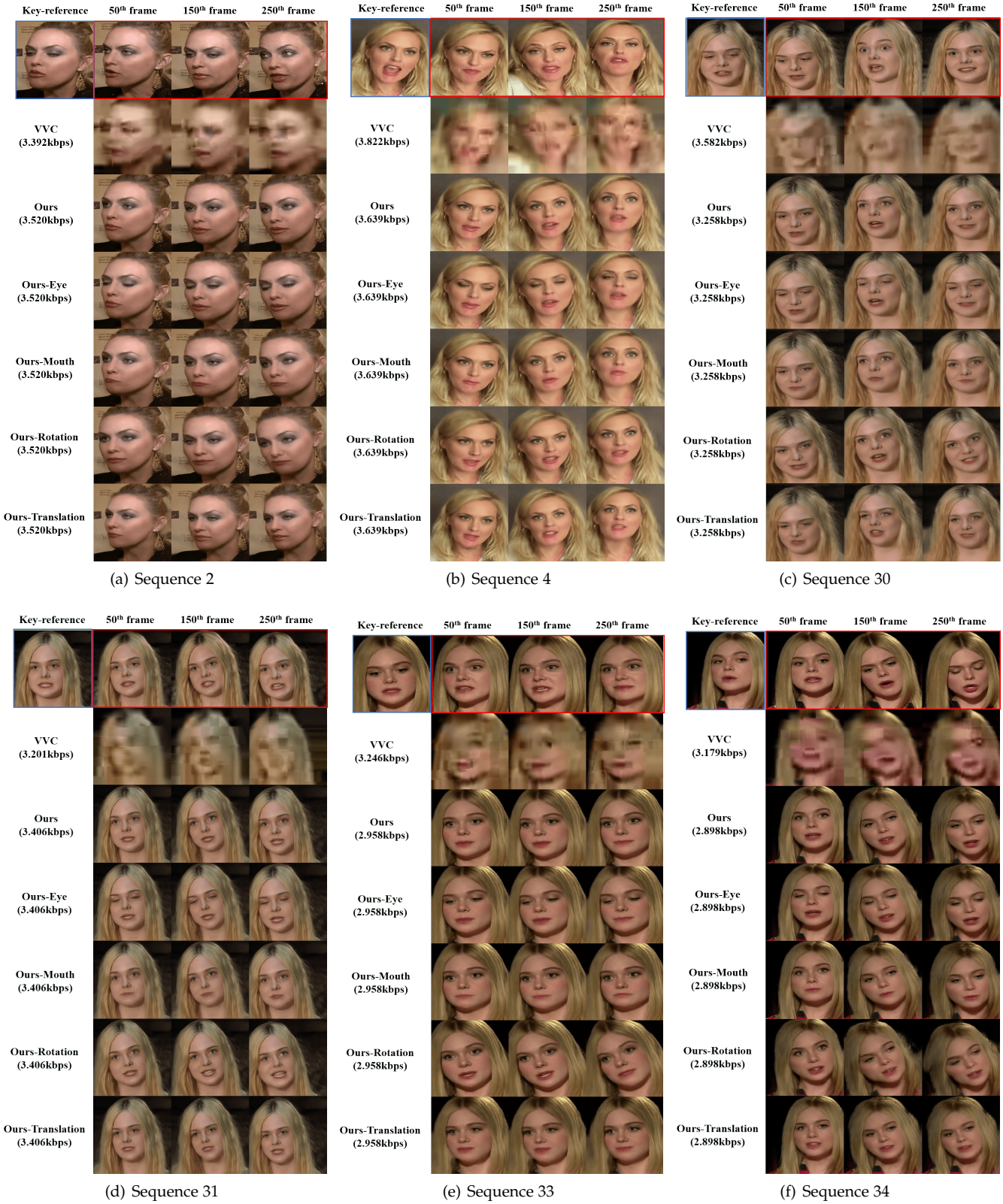
TABLE 1: Bit-rate savings of each video sequence at the resolution of 256×256 in terms of DISTS, LPIPS and FID.

| Seq | Anchor: VVC [8] | | | Anchor: FOMM [9] | | | Anchor: FOMM2.0 [10] | | | Anchor: Face2FaceRHO [13] | | | Anchor: Face_vid2vid [12] | | | Anchor: CFTE [11] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DISTS | LPIPS | FID | DISTS | LPIPS | FID | DISTS | LPIPS | FID | DISTS | LPIPS | FID | DISTS | LPIPS | FID | DISTS | LPIPS | FID |
| 1 | -81.04% | -77.57% | -80.66% | -57.03% | -58.15% | -58.48% | -35.93% | -30.70% | -41.25% | -65.47% | -67.21% | -65.05% | -11.89% | -10.45% | -13.16% | -8.40% | -5.93% | -7.65% |
| 2 | -76.00% | -69.36% | -72.89% | -63.33% | -63.10% | -62.55% | -45.01% | -38.67% | -43.72% | -69.05% | -70.41% | -67.42% | -20.69% | -14.99% | -13.82% | -8.65% | -3.00% | -6.66% |
| 3 | -75.97% | -72.49% | -74.89% | -64.12% | -63.81% | -62.49% | -49.00% | -44.84% | -47.73% | -68.10% | -67.65% | -62.79% | -23.58% | -20.91% | -22.41% | -11.23% | -6.19% | -8.69% |
| 4 | -77.64% | -71.03% | -77.96% | -64.83% | -65.33% | -64.49% | -45.78% | -41.85% | -45.86% | -69.42% | -71.66% | -66.16% | -20.02% | -15.28% | -18.50% | -9.77% | -5.44% | -14.98% |
| 5 | -75.74% | -72.16% | -78.21% | -59.53% | -60.13% | -61.17% | -41.55% | -38.94% | -41.38% | -68.53% | -70.44% | -66.43% | -7.96% | -8.07% | -11.94% | -4.67% | -1.83% | -9.51% |
| 6 | -70.53% | -63.83% | -71.66% | -54.99% | -53.85% | -55.17% | -30.40% | -26.59% | -28.51% | -69.25% | -68.57% | -58.92% | 2.67% | 6.08% | 3.67% | -3.01% | 5.36% | -0.96% |
| 7 | -69.55% | -66.13% | -68.60% | -52.55% | -53.26% | -56.04% | -28.87% | -26.75% | -30.43% | -65.40% | -67.27% | -63.43% | -0.34% | -2.50% | -2.75% | -0.15% | 1.00% | -5.35% |
| 8 | -74.67% | -63.42% | -68.42% | -58.60% | -60.60% | -50.45% | -34.96% | -23.79% | -19.85% | -61.35% | -55.87% | -48.46% | -26.33% | -13.58% | -20.99% | -19.37% | -14.71% | -11.61% |
| 9 | -69.27% | -68.03% | -69.74% | -54.87% | -55.89% | -55.62% | -28.62% | -25.74% | -31.94% | -67.47% | -69.11% | -69.93% | -4.31% | -4.57% | -4.57% | -2.43% | 2.23% | -3.06% |
| 10 | -73.56% | -70.47% | -71.54% | -55.11% | -55.12% | -55.82% | -34.99% | -31.63% | -35.35% | -59.41% | -61.62% | -57.31% | -5.83% | -4.38% | -6.79% | -2.66% | 2.56% | 0.71% |
| 11 | -67.63% | -63.71% | -64.64% | -55.13% | -54.91% | -56.50% | -36.83% | -33.90% | -39.44% | -61.94% | -64.36% | -64.09% | -8.58% | -5.55% | -8.94% | -6.70% | -0.93% | -5.04% |
| 12 | -78.00% | -75.67% | -77.67% | -56.35% | -55.67% | -56.56% | -39.20% | -35.87% | -40.47% | -62.70% | -64.80% | -60.88% | -12.56% | -9.47% | -11.67% | -8.70% | -1.78% | -7.18% |
| 13 | -72.54% | -68.94% | -71.63% | -57.75% | -57.19% | -66.49% | -35.76% | -29.95% | -37.02% | -65.49% | -66.04% | -69.51% | -10.95% | -7.55% | -18.34% | -3.89% | 2.46% | -19.96% |
| 14 | -78.06% | -74.14% | -74.91% | -62.25% | -64.13% | -60.96% | -42.08% | -39.31% | -45.11% | -67.15% | -67.58% | -63.20% | -17.55% | -15.69% | -15.36% | -12.46% | -7.92% | -10.53% |
| 15 | -75.61% | -71.51% | -76.27% | -55.25% | -53.45% | -58.05% | -38.07% | -33.59% | -41.07% | -60.02% | -62.19% | -63.13% | -9.51% | -5.46% | -13.31% | -9.36% | -0.19% | -12.13% |
| 16 | -73.50% | -65.69% | -4.91% | -58.20% | -56.19% | 24.25% | -39.79% | -33.83% | 58.40% | -69.25% | -69.70% | -54.18% | -6.76% | 1.25% | 195.43% | -13.07% | 0.69% | 208.79% |
| 17 | -70.42% | -68.77% | -65.62% | -56.22% | -57.44% | -55.29% | -35.58% | -33.39% | -33.68% | -61.13% | -64.24% | -57.69% | -5.85% | -4.74% | -1.64% | -3.76% | 0.40% | 2.52% |
| 18 | -75.01% | -70.95% | -75.18% | -59.26% | -59.01% | -60.70% | -40.62% | -36.06% | -41.74% | -65.25% | -65.30% | -65.54% | -14.00% | -10.92% | -17.03% | -8.66% | -0.35% | -3.10% |
| 19 | -65.34% | -65.41% | -67.19% | -54.78% | -55.08% | -56.94% | -33.39% | -29.31% | -34.53% | -60.96% | -60.10% | -61.60% | -4.42% | -1.77% | -9.21% | -2.26% | 4.25% | 1.51% |
| 20 | -71.46% | -69.38% | -71.00% | -55.77% | -57.42% | -60.11% | -28.85% | -21.35% | -27.04% | -63.43% | -64.33% | -67.95% | -4.43% | -1.01% | -10.62% | -4.67% | 3.73% | -6.95% |
| 21 | -69.69% | -61.20% | -61.16% | -60.33% | -59.27% | -57.73% | -36.55% | -34.37% | -35.21% | -66.18% | -65.07% | -57.52% | -7.16% | -3.65% | -0.35% | -2.62% | 1.76% | 4.25% |
| 22 | -76.29% | -68.72% | -72.57% | -56.76% | -53.91% | -56.62% | -33.16% | -26.46% | -35.03% | -66.06% | -66.09% | -65.73% | -8.55% | -0.73% | -10.03% | -3.51% | 5.79% | -5.27% |
| 23 | -67.40% | -62.76% | -66.89% | -56.84% | -56.57% | -58.36% | -35.43% | -31.01% | -35.72% | -67.33% | -69.59% | -67.46% | -5.95% | -2.48% | -8.73% | -3.86% | -2.67% | -8.62% |
| 24 | -73.47% | -70.04% | -75.53% | -60.79% | -61.10% | -57.98% | -40.64% | -38.07% | -38.51% | -64.08% | -64.34% | -63.39% | -8.70% | -6.20% | -7.18% | -5.17% | 0.15% | -1.81% |
| 25 | -68.52% | -65.83% | -75.84% | -56.36% | -54.88% | -55.04% | -33.81% | -31.44% | -33.16% | -63.42% | -64.11% | -64.58% | -1.01% | 2.37% | -2.27% | 1.50% | 8.39% | -1.29% |
| 26 | -78.26% | -74.43% | -82.57% | -63.32% | -62.88% | -66.48% | -46.86% | -43.64% | -50.55% | -67.80% | -67.44% | -68.32% | -21.29% | -16.58% | -25.22% | -8.28% | -1.74% | -15.88% |
| 27 | -77.48% | -71.37% | -79.39% | -60.79% | -60.70% | -60.45% | -40.63% | -36.77% | -41.96% | -69.95% | -70.78% | -66.74% | -20.48% | -15.49% | -18.61% | -8.51% | -0.84% | -8.31% |
| 28 | -82.68% | -75.49% | -82.94% | -64.54% | -65.13% | -63.71% | -46.88% | -40.91% | -48.08% | -66.62% | -67.23% | -65.52% | -19.15% | -13.86% | -15.98% | -16.63% | -4.82% | -16.13% |
| 29 | -67.87% | -60.72% | -73.60% | -60.66% | -59.96% | -62.02% | -41.24% | -35.90% | -44.71% | -69.06% | -70.53% | -67.72% | -17.69% | -9.93% | -16.88% | -10.22% | 0.43% | -11.14% |
| 30 | -76.42% | -72.74% | -77.80% | -59.23% | -58.40% | -59.75% | -42.26% | -39.41% | -42.04% | -63.93% | -64.54% | -59.71% | -7.67% | -5.69% | -6.15% | -9.11% | -5.50% | -6.83% |
| 31 | -75.86% | -71.23% | -79.25% | -60.42% | -60.17% | -62.33% | -44.82% | -43.55% | -48.16% | -64.08% | -65.31% | -63.63% | -11.58% | -9.60% | -13.19% | -8.06% | -5.09% | -10.67% |
| 32 | -78.97% | -70.02% | -79.29% | -67.01% | -66.05% | -66.13% | -48.22% | -38.08% | -45.43% | -67.09% | -67.39% | -62.80% | -23.65% | -21.24% | -18.49% | -14.68% | -6.91% | -10.96% |
| 33 | -75.76% | -67.67% | -80.92% | -61.38% | -59.69% | -63.93% | -40.83% | -35.41% | -44.87% | -67.74% | -69.15% | -69.12% | -9.82% | -4.72% | -12.55% | -12.40% | -2.87% | -27.55% |
| 34 | -71.85% | -57.36% | -77.41% | -61.92% | -59.44% | -60.23% | -34.86% | -16.76% | -36.29% | -73.02% | -73.55% | -75.97% | -5.93% | 9.02% | -8.83% | -5.97% | 9.89% | -8.01% |
| 35 | -84.93% | -79.46% | -87.32% | -61.85% | -60.84% | -63.71% | -46.85% | -41.73% | -49.95% | -61.39% | -64.32% | -61.57% | -15.42% | -12.80% | -18.14% | -15.59% | -9.43% | -16.04% |
| 36 | -75.87% | -64.44% | -69.10% | -60.50% | -61.03% | -54.93% | -40.63% | -27.30% | -33.68% | -60.63% | -63.02% | -51.05% | -18.19% | -5.54% | -63.29% | -11.87% | 1.54% | 4.56% |
| 37 | -75.54% | -70.35% | -76.71% | -58.20% | -58.09% | -58.68% | -43.43% | -40.37% | -43.00% | -64.25% | -63.70% | -61.40% | -13.26% | -10.83% | -17.29% | -10.10% | -3.42% | -8.57% |
| 38 | -80.14% | -71.82% | -75.58% | -58.28% | -56.17% | -51.23% | -41.37% | -35.78% | -31.17% | -59.37% | -59.67% | -50.89% | -15.49% | -10.36% | 0.13% | -13.93% | -3.70% | -0.57% |
| 39 | -76.77% | -73.47% | -79.15% | -59.64% | -58.74% | -60.87% | -42.35% | -39.48% | -46.11% | -64.69% | -66.95% | -64.10% | -16.53% | -10.12% | -19.73% | -12.67% | -4.72% | -10.30% |
| 40 | -73.08% | -69.91% | -69.99% | -57.04% | -57.61% | -55.98% | -34.21% | -31.95% | -35.50% | -66.34% | -67.42% | -65.75% | -7.73% | -5.46% | -4.01% | -8.08% | -3.74% | -4.40% |
| 41 | -69.22% | -64.72% | -68.80% | -56.45% | -56.37% | -56.68% | -35.10% | -29.93% | -37.79% | -66.84% | -66.56% | -68.77% | -2.53% | 2.98% | -4.69% | -7.84% | -1.76% | -7.25% |
| 42 | -82.03% | -78.15% | -83.67% | -60.81% | -63.84% | -63.48% | -37.77% | -35.39% | -43.32% | -70.43% | -74.58% | -70.60% | -11.22% | -11.94% | -14.94% | -7.46% | -7.35% | -9.74% |
| 43 | -80.40% | -78.02% | -83.36% | -57.92% | -57.90% | -58.21% | -42.92% | -37.98% | -40.61% | — | — | — | -10.94% | -8.07% | -7.64% | -8.75% | -6.55% | -6.90% |
| 44 | -82.09% | -78.63% | -82.44% | -62.08% | -63.08% | -62.63% | -45.36% | -43.95% | -47.10% | -60.80% | -62.76% | -62.92% | -15.30% | -15.14% | -20.02% | -10.01% | -8.08% | -11.92% |
| 45 | -82.29% | -78.81% | -81.49% | -61.02% | -60.48% | -61.52% | -43.86% | -38.90% | -44.28% | -60.21% | -60.79% | -59.62% | -15.59% | -11.10% | -14.71% | -10.15% | -4.33% | -7.93% |
| 46 | -74.98% | -68.44% | -75.66% | -64.45% | -65.32% | -63.48% | -45.64% | -43.57% | -44.95% | -70.50% | -71.79% | -70.26% | -17.41% | -14.92% | -16.08% | -11.01% | -7.66% | -8.88% |
| 47 | -82.44% | -78.32% | -81.19% | -62.92% | -62.69% | -61.96% | -47.66% | -42.80% | -49.87% | -60.46% | -60.58% | -61.00% | -17.38% | -17.21% | -18.70% | -13.02% | -9.32% | -12.89% |
| 48 | -81.13% | -75.45% | -84.48% | -63.24% | -61.54% | -62.66% | -48.36% | -41.10% | -49.02% | -66.52% | -65.80% | -61.36% | -20.06% | -11.56% | -20.81% | -13.35% | -5.32% | -9.95% |
| 49 | -82.32% | -78.36% | -83.22% | -58.22% | -57.10% | -57.35% | -43.14% | -36.05% | -39.38% | — | — | — | -10.24% | -5.61% | -3.78% | -8.24% | -4.32% | -3.84% |
| 50 | -73.24% | -67.95% | -70.26% | -59.00% | -61.11% | -55.54% | -43.09% | -37.76% | -41.54% | -70.26% | -88.47% | -66.27% | -13.97% | -8.63% | -11.78% | -13.91% | -10.83% | -11.89% |
| **Average** | **-75.37%** | **-70.29%** | **-74.02%** | **-59.36%** | **-59.20%** | **-57.79%** | **-39.86%** | **-35.04%** | **-38.09%** | **-65.41%** | **-66.67%** | **-63.49%** | **-12.06%** | **-8.09%** | **-8.84%** | **-8.55%** | **-2.37%** | **-3.49%** |


(a) Sequence 2 (b) Sequence 4 (c) Sequence 10


(d) Sequence 31 (e) Sequence 33 (f) Sequence 34

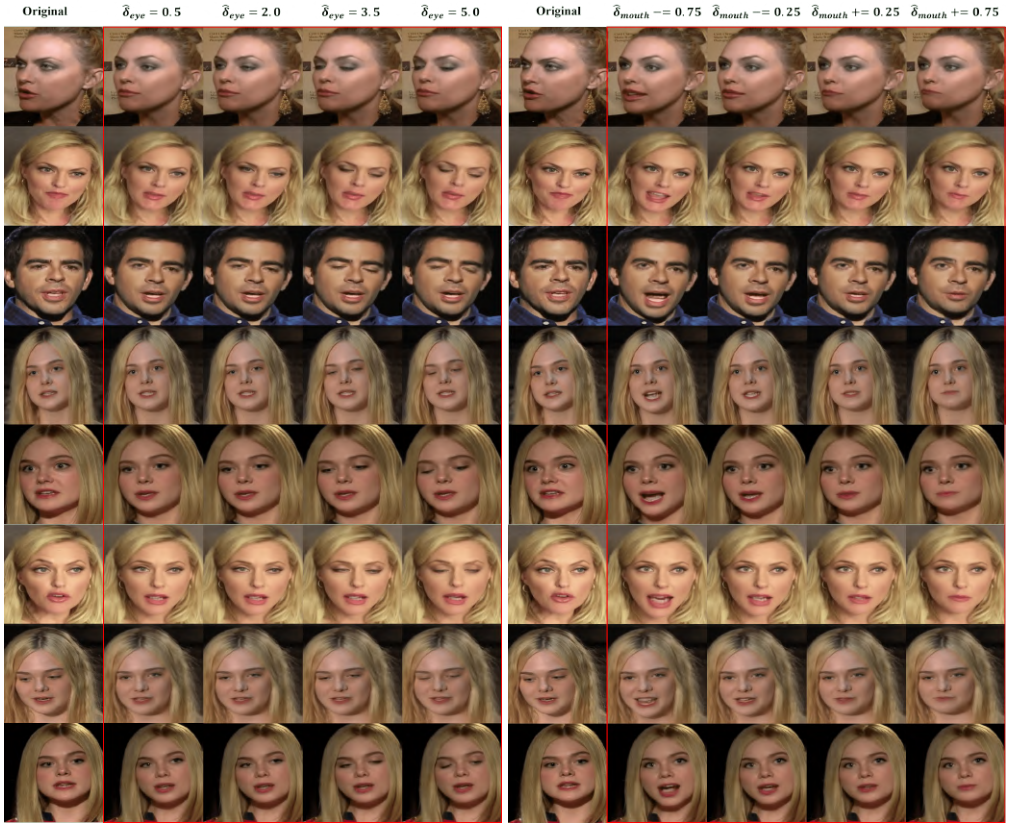
Fig. 2: Visual quality comparisons among VVC [8], FOMM [9], FOMM2.0 [10], Face2FaceRHO [13], Face_vid2vid [12], CFTE [11] and Ours at the similar bit rate. The values on the left represent coding bits and DISTS value, where lower DISTS value indicates better quality. More video examples can be found in project page.



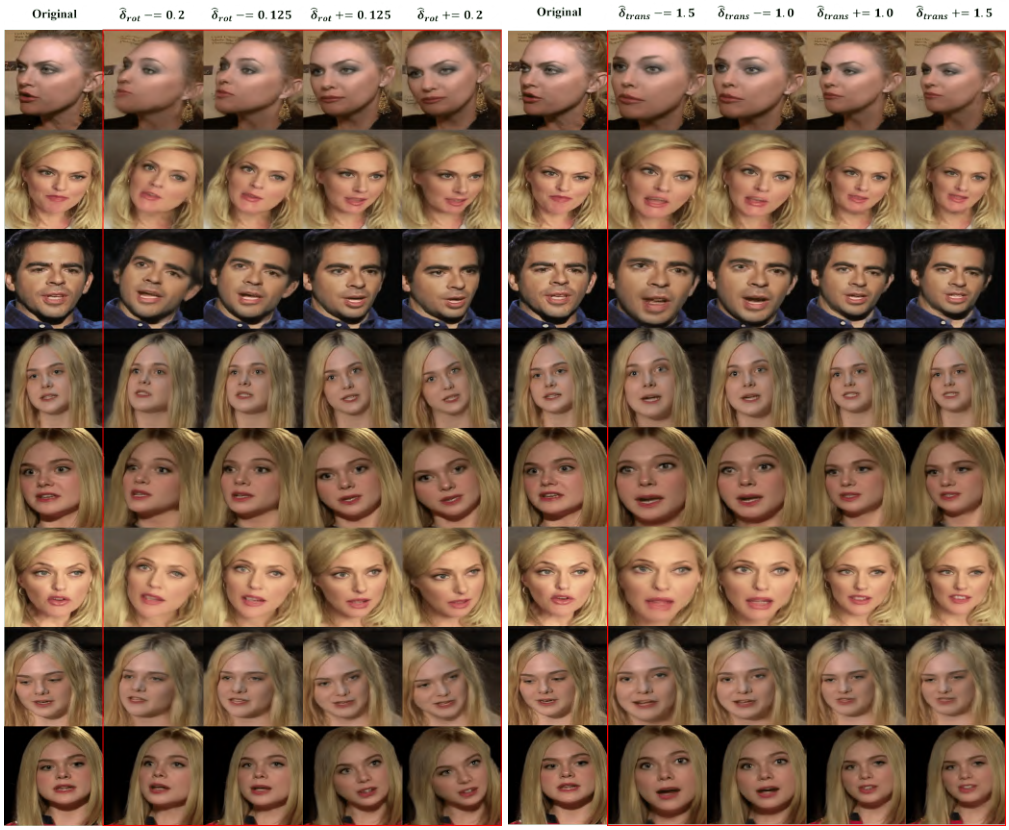

(a) Sequence 2 (b) Sequence 4 (c) Sequence 30

(d) Sequence 31 (e) Sequence 33 (f) Sequence 34

Fig. 3: Examples on animating virtual character reference with facial semantics. More video examples can be found in project page.

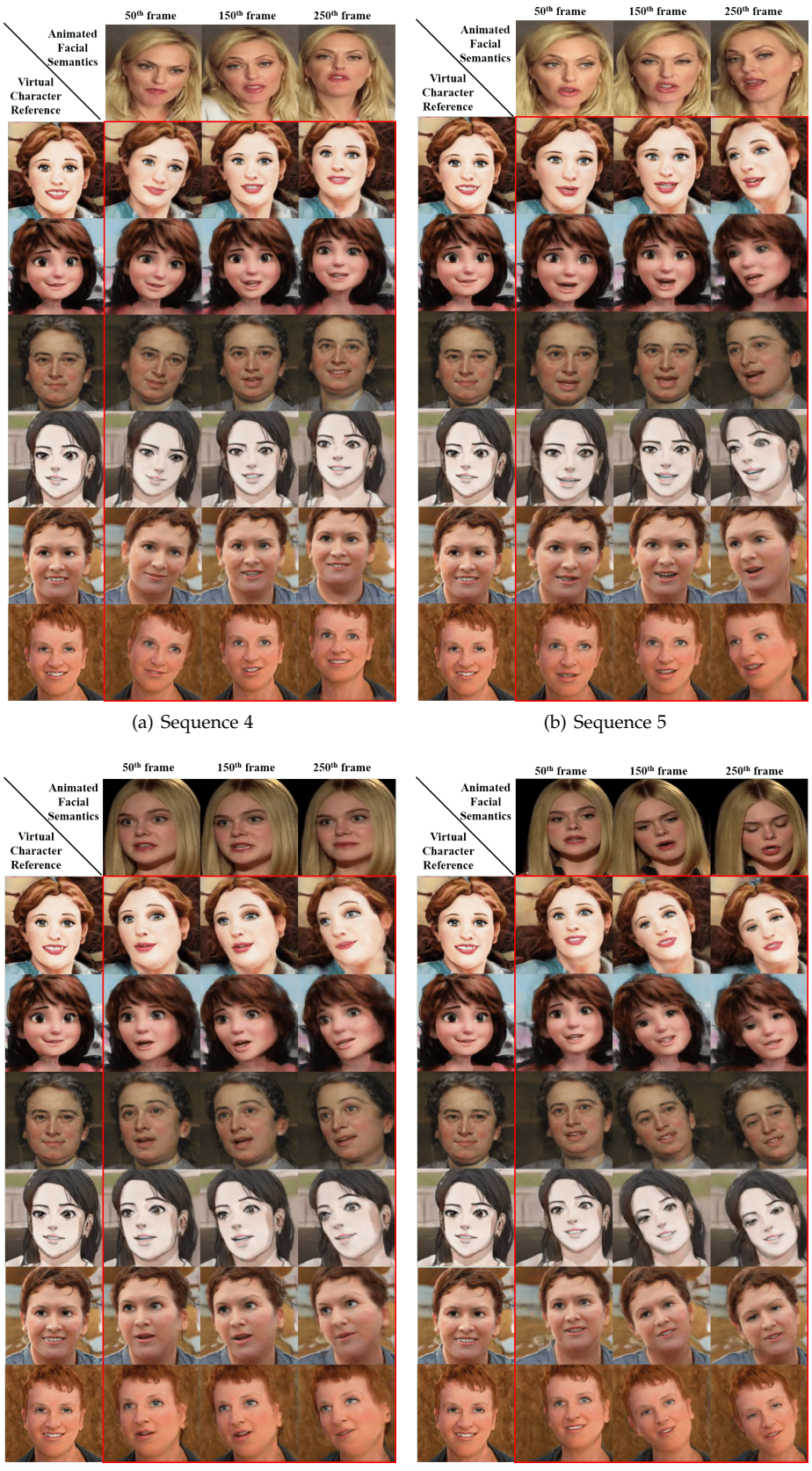



(a) Eye Motion $\hat{\delta}_{eye}$ (b) Mouth Motion $\hat{\delta}_{mouth}$




(c) Head Rotation $\hat{\delta}_{rot}$ (d) Head Translation $\hat{\delta}_{trans}$

Fig. 4: Examples on face interactive coding in terms of eye motion, mouth motion, head rotation and head translation. Different columns represent different interactive degree of talking face frames. More video examples can be found in project page.




(a) Sequence 4 (b) Sequence 5

(c) Sequence 33 (d) Sequence 34

Fig. 5: Examples on animating virtual character reference with facial semantics. More video examples can be found in project page.