Total-Editing: Head Avatar with Editable Appearance, Motion, and Lighting

Supplementary Material

A. Training Scheme and Objective Functions

We begin by pre-training our lightmap estimator using synthetic data with ground truths. The objective function is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\mathbf{S}} + \mathcal{L}_{\mathbf{N}}, \tag{22}$$

where $\mathcal{L}_{\mathbf{S}}$ is an L1 loss comparing the predicted and ground truth lightmaps $\mathbf{S}_{\text{d}}, \{\mathbf{S}_{\text{s}}(n)\}$, $\mathcal{L}_{\mathbf{N}}$ is a cosine similarity loss between predicted and ground truth normal $\mathbf{N}$. After pre-training, we detach the normal decoder $\mathcal{E}_{\text{nor}}$ and integrate the rest of the lightmap estimator with Total-Editing. The entire network is then trained end-to-end using both real and synthetic data. During training, we randomly sample appearance source $\mathbf{I}_{\text{app}}$ and motion source $\mathbf{I}_{\text{mot}}$ of one subject, with the editing target $\mathbf{I}^*$ equal to the motion source $\mathbf{I}_{\text{mot}}$. As for the lighting source, we use the HDR environment map of $\mathbf{I}_{\text{mot}}$ for synthetic data and another random frame from the same video clip for real data. In this phase, our reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\mathbf{a}} + \mathcal{L}_{\mathbf{n}} + \mathcal{L}_{\mathbf{S}}, \tag{23}$$

where $\mathcal{L}_1$, $\mathcal{L}_{\text{LPIPS}}$, $\mathcal{L}_{\text{id}}$ are pixel-wise L1, perceptual difference [80], and negative cosine similarity of face recognition features [14] between the editing result $\hat{\mathbf{I}}$ and the target $\mathbf{I}^*$, $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\mathbf{a}}$ are L1 losses for the rendered foreground mask and albedo, $\mathcal{L}_{\mathbf{n}}$ is a cosine similarity loss for the rendered normal, and $\mathcal{L}_{\mathbf{S}}$ is the L1 loss for estimated lightmaps. Note that $\mathcal{L}_{\mathbf{a}}$ and $\mathcal{L}_{\mathbf{S}}$ are used only for synthetic data, while $\mathcal{L}_{\mathbf{n}}$ is also applied to real data with Sapiens [33] pseudo ground truths. Further, we introduce regularization

$$\mathcal{L}_{\text{reg}} = \mathcal{R}_{\text{TV}} + \mathcal{R}_{\delta} + \mathcal{R}_{\mathbf{n}}, \tag{24}$$

where $\mathcal{R}_{\text{TV}}$ is the total variation loss to promote spatial smoothness, $\mathcal{R}_{\delta}$ is a L1 regularization for residual color $\delta\mathbf{c}$ which constraints it from dominating the render, and

$$\mathcal{R}_{\mathbf{n}} = \left\| 1 - \mathbf{n} \cdot \left(-\frac{\nabla\sigma(\mathbf{p})}{\|\nabla\sigma(\mathbf{p})\|_2} \right) \right\|_1 \tag{25}$$

regularizes normal $\mathbf{n}$ to align with the unit negative gradient of density σ. We also apply an adversarial loss $\mathcal{L}_{\text{adv}}$ with a dual discriminator [8]. Finally, the training objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{adv}}. \tag{26}$$

B. More Qualitative Results

We present additional face reenactment results on the VFHQ [71] and HDTF [81] datasets in Figs. 11 to 14, demonstrating the effectiveness of Total-Editing in both motion and lighting control. Further, we explore two downstream applications. As shown in Fig. 15, Total-Editing enables relighting of animated portraits using HDR environment maps, producing a background replacement effect. In Fig. 16, Total-Editing can leverage arbitrary portrait images as lighting sources, vividly transferring the illumination effect from one portrait to another.

Figure 10. **Samples from one identity of the Lumos [75] dataset.** We refer to each column as a unique subject, since they have different appearances and accessories.

C. Dataset Details

Our synthetic dataset is designed to advance research in general portrait editing, offering two primary subsets: a multi-view subset and a video-like subset.

- The multi-view subset comprises 50K subjects, each captured in two distinct environments and viewed from 10 camera angles. This subset provides extensive data for analyzing objects from diverse perspectives and ensuring multi-view consistency. Samples are shown in Fig. 17.
- The video-like subset includes 10K subjects, each rendered across 10 environments with varied poses and expressions, making it well-suited for studying motion and temporal changes. Samples are shown in Fig. 18.

With diverse samples demonstrated in Fig. 19, our synthetic dataset consists of 50K subjects and 2M images. It is enriched with ground truth albedo, normal, depth, UV maps, segmentation masks, and HDR environment maps. This dataset addresses critical limitations compared to the existing synthetic datasets. For example, as shown in Fig. 10, the Lumos dataset [75] captures each subject only from one view, limiting it to tasks like portrait relighting. In contrast, our dataset incorporates multiple viewpoints and subject movements, better simulating real-world spatiotemporal variations. These improvements make our dataset more versatile and effective for downstream applications requir-

(a) Appearance (b) Motion (c) GPA (d) GPA + PR (e) R3DP (f) R3DP + PR (g) P4Dv2 (h) P4Dv2 + PR (i) Ours

Figure 11. **Additional cross-reenactment results on the VFHQ dataset.**

ing spatial and/or temporal coherence.

D. Evaluation Details

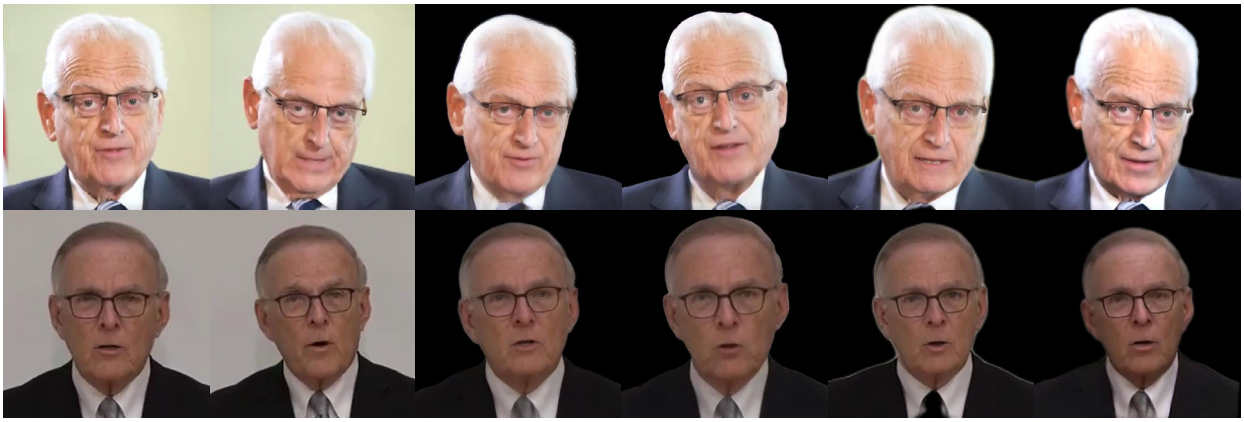
In Tab. 1, we exclude both generated and ground truth backgrounds from metric calculations, focusing solely on the quality of the portrait regions. Since the relighting method proposed by Cai et al. [6] requires additional cropping for input images, we recompose the outputs with original inputs in Fig. 7 for visual consistency. During quantitative comparisons in Tab. 1, we only consider the valid areas after cropping. In Fig. 9, the backgrounds of 2^{nd} to 3^{rd} columns are synthesized by inpainting the lighting source portrait using [62], while those in the 4^{th} to 5^{th} columns are rendered from the corresponding HDR environment maps.

(a) Appearance (b) Motion (c) GPA (d) R3DP (e) P4Dv2 (f) Ours

Figure 12. **Cross-reenactment results on the HTDF dataset.**

(a) Appearance (b) Motion (c) GPA (d) R3DP (e) P4Dv2 (f) Ours

Figure 13. **Cross-reenactment results on the HTDF dataset (continued).**

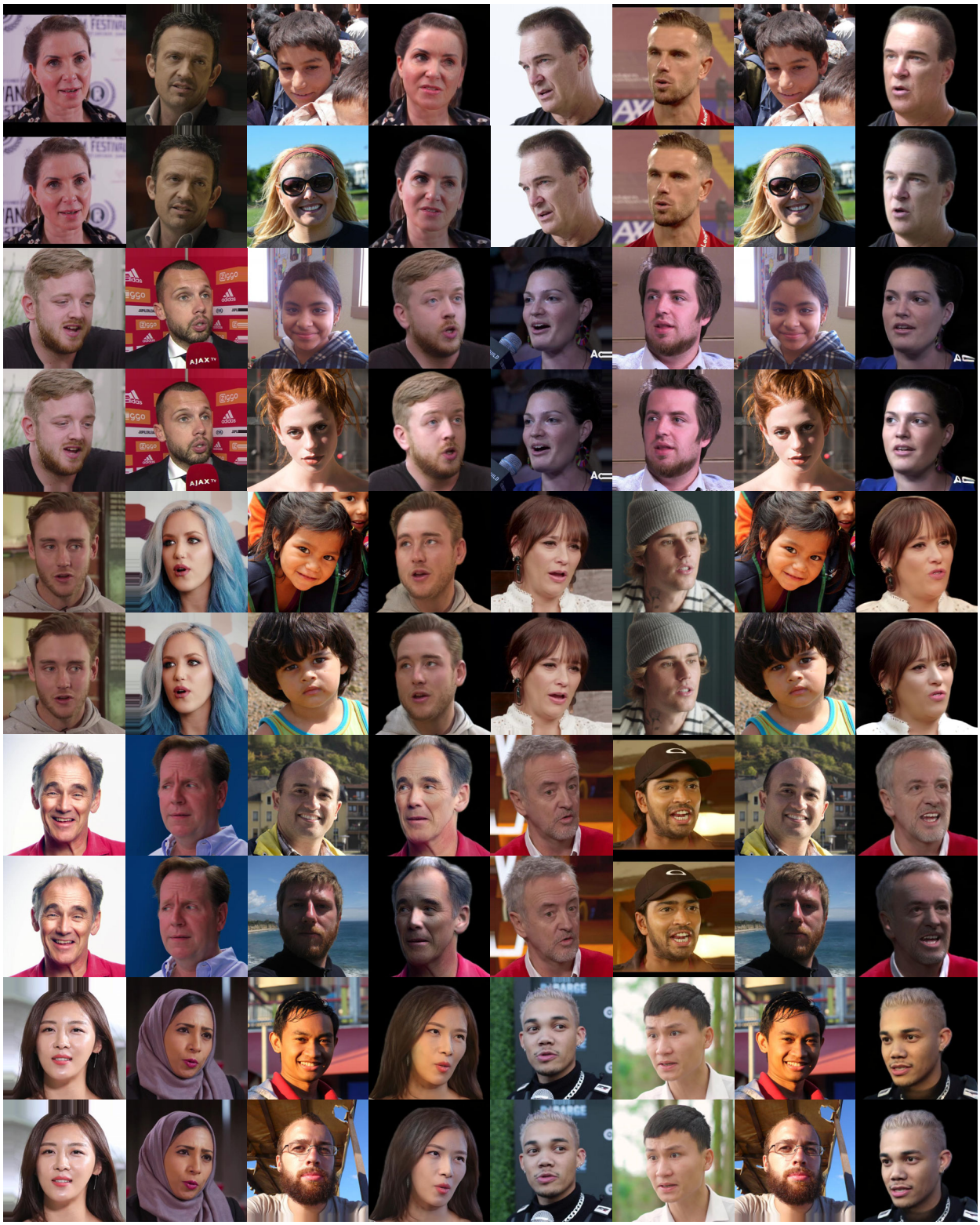


(a) Appearance (b) Motion (c) GPA (d) R3DP (e) P4Dv2 (f) Ours

Figure 14. **Self-reenactment results on the HTDF dataset.**

(a) Appearance (b) Motion (c) Lighting (d) Ours (e) Appearance (f) Motion (g) Lighting (h) Ours

Figure 15. **Cross-reenactment results on the VFHQ dataset with HDR environment maps as lighting sources.**



(a) Appearance (b) Motion (c) Lighting (d) Ours (e) Appearance (f) Motion (g) Lighting (h) Ours

Figure 16. **Cross-reenactment results on the VFHQ dataset with portrait images as lighting sources.**

Figure 17. **Multi-view subset of our synthetic data.** It incorporates 50K subjects. Each is rendered in 2 environments with 10 views.

Figure 18. **Video subset of our synthetic data.** It includes 10K subjects. Each is rendered in 10 environments with 10 poses/expressions.

Figure 19. **More subjects in our synthetic data.** Subjects are with randomized poses, expressions, hairstyles, skin types, accessories, etc.