Supplementary Material of CelebV-HQ: A Large-Scale Video Facial Attributes Dataset

Hao Zhu[1*][0000−0003−2155−1488], Wayne Wu[1*✉][0000−0002−1364−8151], Wentao Zhu[2][0000−0002−5483−0259], Liming Jiang[3][0000−0001−8109−5598], Siwei Tang[1][0000−0002−3105−551X], Li Zhang[1][0000−0002−8714−2137], Ziwei Liu[3][0000−0002−4220−5958], and Chen Change Loy[3][0000−0001−5345−1591]

[1] SenseTime Research
[2] Peking University
[3] S-Lab, Nanyang Technological University

1 Data Pre-processing

We provide an illustration of data pre-processing as shown in Fig. A1. First, we detect facial landmarks [15] from the video, and use these to extract the bounding boxes. Faces smaller than 450 pixels are filtered out. Then, we check whether the adjacent frames belong to the same person based on the motion [2] and identity [6]. If not, we will split the video into different clips. Next, given a sequence of bounding boxes, we calculate their minimum bounding rectangle. To reduce data loss, we expand the bounding rectangle smaller than 512^2 to this size, and use the bounding box to crop the original video. Finally, only clips longer than 3 seconds are kept.

2 Additional Statistic Comparisons

2.1 Comparison of appearance attribute statistics with CelebA-HQ

As shown in Fig. A2, CelebV-HQ has a similar distribution to CelebA-HQ, and the distribution of most appearance attributes is close to that of CelebA-HQ. This indicates that there is no significant deviation in the distribution of CelebV-HQ.

2.2 Face Shape Ratio.

The distribution of face shape ratio indicates the diversity of the dataset in terms of face types. Therefore, a simple analysis of face shape is proposed, where we calculate the ratio of a face using key points [15] as shown in Fig. A3 (a). The distance from the left and right of the cheeks to the nose is recorded as the width,

* Equal contribution.
✉ Corresponding author (wuwenyan0503@gmail.com).

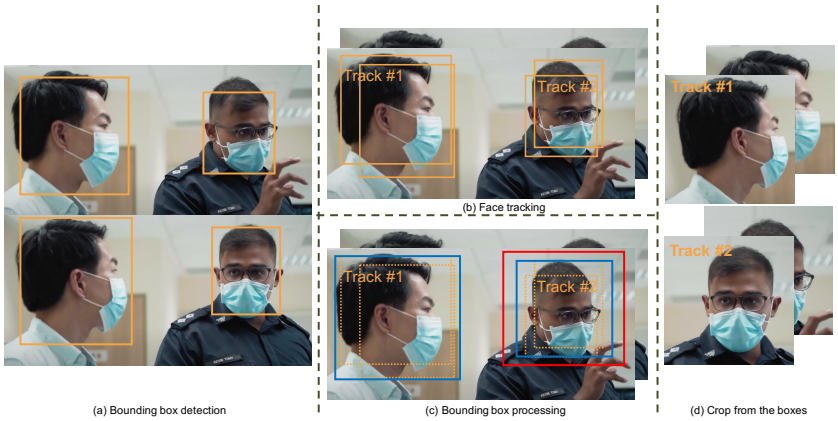


Fig. A1. Pipeline of data pre-process (a) We start from the bounding box detection for each frame. (b) A tracking framework [2] is introduced to track different identities. (c) Given bounding box sequences (dotted orange boxes), we calculate their minimum bounding rectangles (blue box). If bounding rectangles smaller than 512×512, we expand it to this size (red box). (d) Finally, the videos are cropped using the bounding rectangles (blue/red boxes).

and the distance from the highest point of the cheeks to the chin is recorded as the height. The width-to-height ratio is used as the definition of the face shape ratio. As shown in Fig. A3 (b), CelebV-HQ has a more uniform distribution, which indicates that the samples in it have diverse face types.

2.3 Comparison of clip duration statistics with Vox

As reported in Fig. A4, the clip time distribution is shorter compared to Vox [13] for ensuring video consistency and annotation accuracy. Also, the videos in CelebV-HQ are all less than 20s, this is because we truncate all the videos at 20s to avoid the attributes changing in the long video.

2.4 Action Unit Analysis.

Facial Action Units (AUs) are the basic actions of a muscle or muscle group, and we use [7] to detect AUs. The dataset is analyzed in both muscle movement richness and naturalness. Fig. A5 (a) shows that CelebV-HQ is more uniformly distributed over different AU values that represents action strength. The main reason is that videos in VoxCeleb2 [5] are mainly talking videos, while CelebV-HQ consists of more types of facial actions. Meanwhile, the smoothness is measured by log dimensionless jerk [1]. As shown in Fig. A5 (b), CelebV-HQ is smoother than VoxCeleb2 [5], as we highlight with the "Mean value line". More AU results are presented in Appendix 2.4.

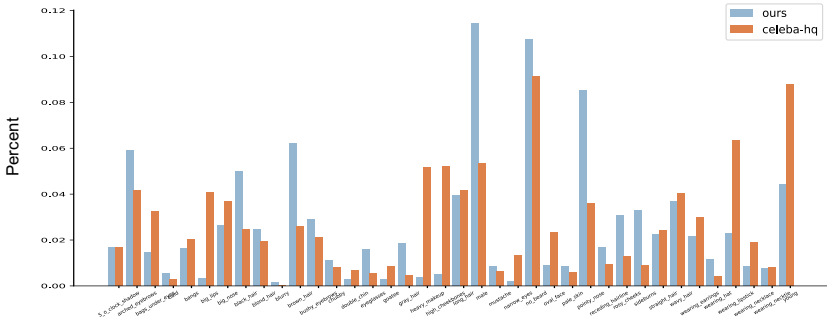


Fig. A2. Comparison of appearance attribute statistics with CelebA-HQ [10]. Please zoom in for more details.

We also provide additional action units (AUs) distributions, as shown in Fig.A6. Fig. A6 (c) and (f) show the locations represented by the different AUs. In Fig. A6 (a) and (d), we can see that the action of CelebV-HQ is smoother than VoxCeleb2 [5]. Meanwhile, Fig. A6 (b) and (e) suggest that CelebV-HQ is more evenly distributed at different AU values.

3 Additional Experiments

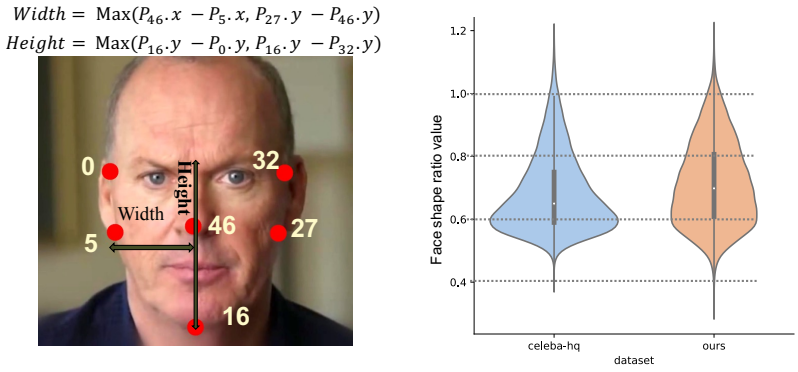
3.1 FVD/FID Setting Details

We leverage FID[4] [8] and FVD[5] [14] to assess the image and video quality of the video generation and editing models. As both metrics are sensitive to the amount of data in the test set, we first select 2048 videos randomly as our test set. All videos in the test set are used as the "real" part in the metric experiments. For the unconditional generation, we also randomly generate 2048 videos as the "fake" part. For the editing of video facial attribution, we generate corresponding fake results for each real video, yielding 2048 fake videos as well. To provide enough images for FID testing, we sample 4 frames from each video. In total, we have 8192 images for the real data and fake data respectively. For the FVD, we use all the real and generated videos.

3.2 Additional Video Facial Attribute Editing Results

To demonstrate the practical value of our dataset for facial attributes editing in low-level appearance attribute. We additional select "Brown Hair" attributes for StarGAN-v2 [4], as well as "Eyeglasses" attributes for MUNIT [9]. The additional results are reported in Table A1 and Fig. A7. By simply adding a temporal

[4] https://github.com/mseitzer/pytorch-fid
[5] https://github.com/sihyun-yu/digan/tree/master/src/metrics



(a) Face shape ratio definition (b) Face shape ratio distribution

Fig. A3. Distributions of head pose and face shape ratio compared with CelebA-HQ [10]. CelebV-HQ contains more diverse head pose and face shape ratio distribution.

Table A1. Quantitative results of video facial attribute editing. We evaluate two video facial editing baselines. The "Video" version achieves lower FVD scores and comparable FID performance than "Original". "↓" means a lower value is better.

Metrics	StarGAN-v2 (Brown Hair)				MUNIT (Eyeglasses)	
	Original		Video		Original	Video
	Reference	Label	Reference	Label		
FVD (↓)	323.71	244.58	**295.74**	**232.63**	204.12	**158.87**
FID (↓)	**77.26**	**64.82**	89.07	69.68	**30.65**	31.23

regularization term, we improve the results of StarGAN-v2 [4] and MUNIT [9] in terms of realism and coherence. Note that the temporal regularization is enabled by CelebV-HQ which contains rich annotations and facial dynamics.

3.3 Experiment on labeled Vox

We labeled Vox dataset [13] using an open-source algorithm [6]. As reported in Table A2, models trained on CelebV-HQ yields better performance. Experiment verified algorithmically labeling existing dataset is not suitable substitutes for CelebV-HQ.

4 Complete Attributes List

4.1 Attribute Selection Details.

For appearance attributes, we derive most of the classes from CelebA [11]. However, we find that three common attributes (, "long hair", "sunglasses" and

[6] https://github.com/ewrfcas/face_attribute_classification_pytorch

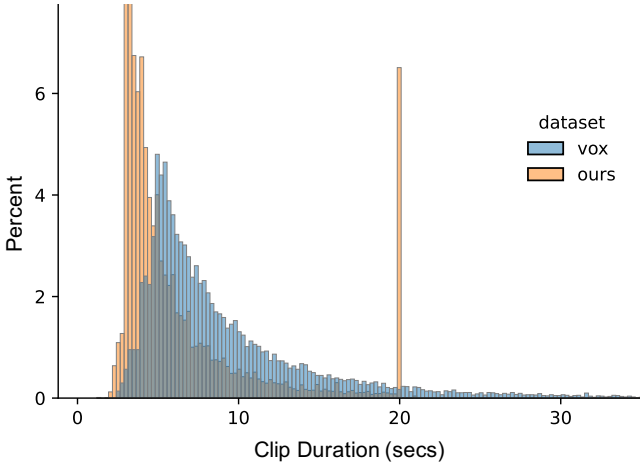
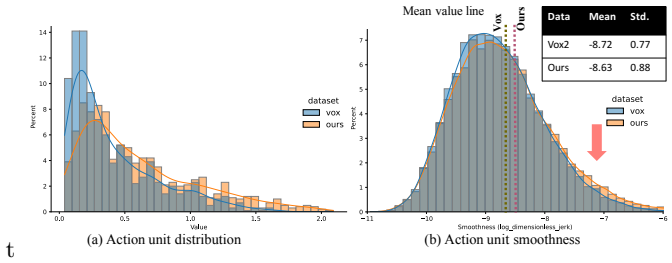


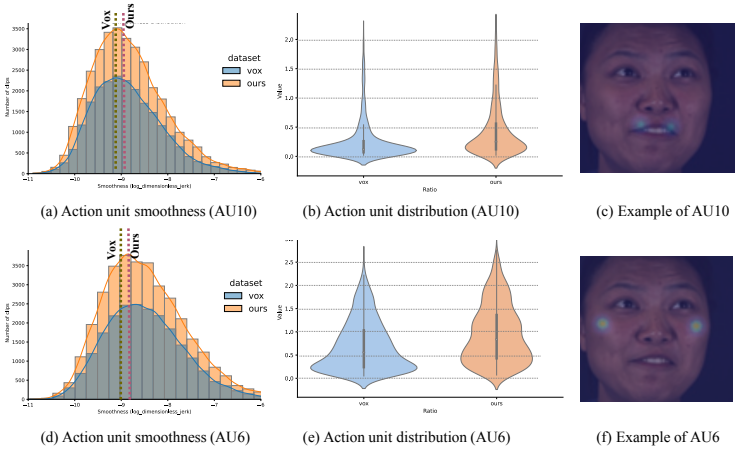
Fig. A4. Comparsion of clip duration statistics with Vox [13].



t

Fig. A5. Distribution and smoothness of action units. We evaluate the distribution (a) and smoothness (b) of action units.

"wearing a mask") in real-world videos are not defined in CelebA [11]. We add these three attributes to the appearance attributes as well. Meanwhile, some action-related attributes, such as "smiling" and "mouth slightly open", have been removed. This process yields 40 appearance attributes in total. For action attributes, inspired by Kinetics-700 [3], we select the face-related actions from its classes and add other facial actions from Internet tags to ensure that the final 35 attributes could cover common facial actions. For emotion attributes, we follow the 8 emotions designed in RAVESS [12], including neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise. Note that the appearance and action attributes are all multi-label as the classes are not mutually exclusive, while emotion attributes are designed to be single-label.



(a) Action unit smoothness (AU10) (b) Action unit distribution (AU10) (c) Example of AU10

(d) Action unit smoothness (AU6) (e) Action unit distribution (AU6) (f) Example of AU6

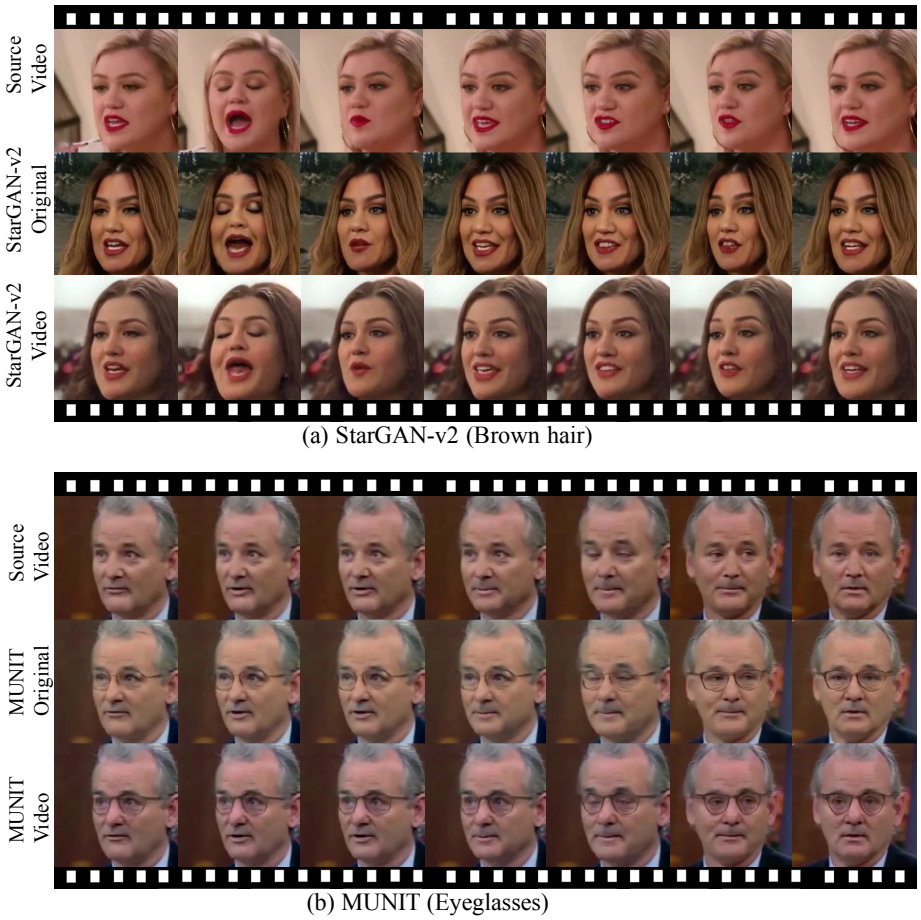
Fig. A6. Distributions of different AUs.

Table A2. Quantitative results of video facial attribute editing on labeled Vox dataset.

Metrics	StarGAN-v2 (Gender)				StarGAN-v2 (Brown hair)			
	Vox-labeled		CelebV-HQ (Ours)		Vox-labeled		CelebV-HQ (Ours)	
	Reference	Label	Reference	Label	Reference	Label	Reference	Label
FVD (↓)	568.79	629.09	**262.01**	**189.04**	542.88	500.77	**295.74**	**232.63**
FID (↓)	104.00	85.14	**82.99**	**55.73**	99.57	131.18	**89.07**	**69.68**

Table A3. Complete attribute list. CelebV-HQ contains 83 annotations, including 40 appearance attributes, 35 action attributes, and 8 emotion attributes.

(a) Appearance Attribute							
blurry	male	young	chubby	pale_skin	rosy_cheeks	oval_face	receding hairline
bald	bangs	black_hair	blond_hair	gray_hair	brown_hair	straight hair	wavy_hair
long_hair	arched eyebrows	bushy eyebrows	bags_under_eyes	eyeglasses	sunglasses	narrow_eyes	big_nose
pointy_nose	high cheekbones	big_lips	double_chin	no_beard	5_o_clock shadow	goatee	sideburns
mustache	heavy makeup	wearing earrings	wearing_hat	wearing lipstick	wearing necklace	wearing necktie	wearing mask
(b) Action Attributes							
blow	chew	close_eyes	cough	cry	drink	eat	frown
gaze	glare	head_wagging	kiss	laugh	listen_to_music	look_around	make_a_face
nod	play_instrument	read	shake_head	shout	sign	sing	sleep
smile	smoke	sneeze	sneer	sniff	talk	turn	weep
whisper	wink	yawn					
(c) Emotion Attributes							
neutral	anger	contempt	disgust	fear	happy	sadness	surprise



(a) StarGAN-v2 (Brown hair)

(b) MUNIT (Eyeglasses)

Fig. A7. Qualitative results of video facial attribute editing. In (a), we edit the attribute *brown hair* with StarGAN-v2 [4]. In (b), we edit the attribute *eyeglasses* with MUNIT [9]. The "video" versions denote the models trained with our temporal constraint.

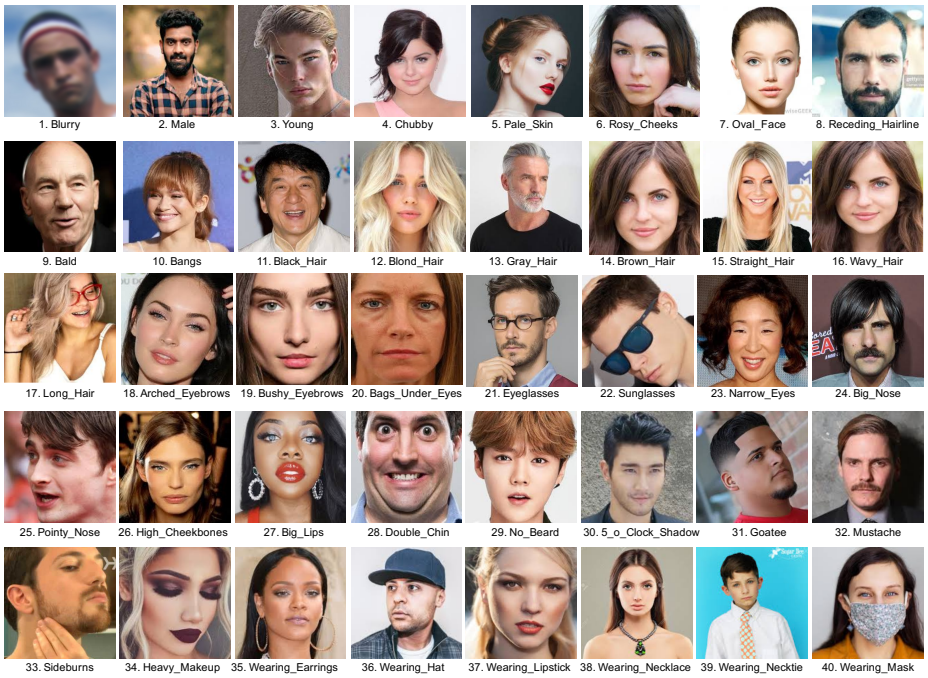


Fig. A8. Examples of appearance attributes.

References

1. Balasubramanian, S., Melendez-Calderon, A., Burdet, E.: A robust and sensitive metric for quantifying movement smoothness. IEEE TBE **59**, 2126–2136 (2011)
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP) (2016)
3. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint **arxiv:1907.06987** (2019)
4. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR (2020)
5. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
6. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
7. Fan, Y., Lam, J., Li, V.: Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In: AAAI (2020)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
9. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
10. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
11. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
12. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS One **13**, e0196391 (2018)
13. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
14. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint **arxiv:1812.01717** (2018)
15. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR (2018)