CelebV-Text: A Large-Scale Facial Text-Video Dataset Supplementary Material

Jianhui Yu[1*] Hao Zhu[2*] Liming Jiang[3] Chen Change Loy[3] Weidong Cai[1] Wayne Wu[4]
[1]University of Sydney [2]SenseTime Research [3]S-Lab, Nanyang Technological University [4]Shanghai AI Laboratory
jianhui.yu@sydney.edu.au haozhu96@gmail.com {liming002,ccloy}@ntu.edu.sg
tom.cai@sydney.edu.au wuwenyan0503@gmail.com

A. Details of Attribute Designs

A.1. Complete Attribute Lists

The complete list of all the attributes is reported in Table A1.

A.2. Grouped Attribute Details

In the main paper, in order to better present the distributions, we divide 40 appearance attributes into facial features, elementary, beard type, hairstyle, and accessories.
a. Facial features: double chin, pale skin, high cheekbones, chubby, oval face, bushy eyebrows, bags under eyes, narrow eyes, heavy makeup, arched eyebrows, pointy nose, big nose, big lips.
b. Elementary: young, male, blurry.
c. Beard type: 5 o'clock shadow, no beard, goatee, sideburns, mustache.
d. Hairstyle: blond hair, gray hair, brown hair, black hair, wavy hair, receding hairline, bangs, straight hair, bald.
e. Accessories: wearing earrings, wearing hat, wearing necktie, wearing necklace, eyeglasses, wearing lipstick

Moreover, all 37 actions are split into Head, Eyes, Interaction, Feeling, and Daily groups.
a. Head: talk, head wagging, look around, turn, shake head, nod.
b. Eyes: blink, wink, squint, close eyes
c. Interaction: drink, sing, eat, smoke, listen to music, play instrument, read, kiss , whisper.
d. Feeling: sneer, sigh, frown, weep, cry, smile, glare, gaze, laugh, shout.
e. Daily: yawn, sneeze, cough, sleep, make a face, smoke, blow, sniff, chew.

A.3. More Distributions

To show the reasonable distribution of CelebV-Text, we first compare the video length duration of our collected videos with CelebV-HQ [18] in Figure A1, where video duration in CelebV-Text is longer than CelebV-HQ. Moreover, the average time duration of CelebV-Text is 14.34s, which is twice more than that of CelebV-HQ of 6.68s. We then present the detailed distributions of general appearances, hair colors, actions and emotions following CelebV-HQ [18] in Figure A2. More distributions of detailed appearances, color temperatures, and brightness are shown in Figure A3. Finally, we compare with CelebV-HQ [18] in more general attributes such as age and ethnicity. Since age and ethnicity labels are not manually annotated, we estimate these two attributes using an off-the-shelf facial attribute analysis framework[1]. As illustrated in Figure A4, CelebV-Text achieves the distributions close to those of CelebV-HQ.

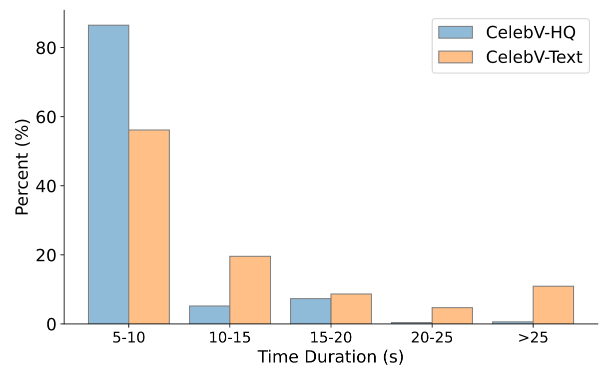


Figure A1. Video time duration of CelebV-Text compared with CelebV-HQ [18].

A.4. Selected Algorithms

For effective and accurate annotation algorithms, we labeled CelebV-Text using an open-source algorithm[2]. We follow [7] for light color temperature and we simplify the light intensity calculation by using perceived brightness [2]. We follow [12] for 8 emotion classification, where emotion label is given for each video frame. We further apply sliding window smoothing algorithm [13] on the temporal domain to smooth the distribution of emotion along time. All automatically annotated labels are further reviewed by our human annotators.

*Equal contribution.

[1]https://github.com/serengil/deepface
[2]https://github.com/ewrfcas/face_attribute_classification_pytorch

Table A1. **Complete attribute list.** CelebV-Text contains both static and dynamic attributes, including 40 general appearances, 5 detailed appearances, 6 light conditions, 37 actions, 8 emotions, and 6 light directions.

Static Attributes					
(a) General Appearance					
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	attractive	arched eyebrows
bushy eyebrows	bags_under_eyes	eyeglasses	mouth_slightly_open	smiling	big_nose
pointy_nose	high cheeks	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	narrow_eyes		
(b) Detailed Appearance					
Mole	freckle	one_eyed	scar	dimple	
(c) Light Conditions					
dark	normal	bright	warm white	cool white	daylight
Dynamic Attributes					
(a) Action					
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	sigh	sing	sleep
smile	smoke	sneeze	sniff	sneer	talk
turn	weep	whisper	win	yawn	blink
squint					
(b) Emotion					
Neutral	Happy	Sad	Anger	Fear	Surprise
Contempt	Disgust				
(c) Light Directions					
front	left_45	right_45	left_90	right_90	back

B. Template Designs

For template design, we first employ trained probabilistic natural language English parsers [4, 5] to parse the natural language inputs provided by out annotators and get parsing tree banks that appear the most. Then we modify the parsing to reversely generate descriptions that are near natural languages. We further choose probabilistic context free grammars (PCFG) to increase the diversity of the generated sentences. One PCFG template used to generate language descriptions for our general face appearance is shown in Table A2. Note that all terminal symbols are bold, and terminal symbol with underlines are dependent on the annotated results. Specifically, **gender_related_attributes** is related the gender, which is a unique value. **personal_noun** is also gender related and can be considered as a list where only one single option is picked (*i.e.*, man, woman, male, female). **wear_related_attributes** contains a list of general attributes related to wearing (*i.e.*, heavy makeup, earrings, hat, lipstick, necklace, necktie, eyeglasses). **is_related_attributes** contains a list of general attributes such as bald, young, blurry. **has_related_attributes** contains 5 o'clock shadow, bags under eyes, arched eyebrows, and so on. Please refer to our GitHub for all designed templates. After obtaining the full sentence, we further use NLTK [3] for synonym replacement to increase our generation diversity.

C. Results of n-grams

We further compare more unique n-grams among MM-Vox [6], CelebV-HQ [18], and CelebV-Text in Table A3. The improvement of our CelebV-Text over MM-Vox [6], CelebV-HQ [18] is quite obvious, which indicates CelebV-Text presents more diverse descriptions.

D. Additional Experiments

D.1. FVD/FID/CLIPSIM Settings

We leverage FVD[3] [16], FID[4] [8], and CLIPSIM[5] [6] to assess the video temporal consistency, individual frame quality, and relevance between the generated video and input text. As all metrics are sensitive to data scale during testing, we first randomly select 2,048 videos from the test data as our "test set", which are used as the "real" part in our metric experiments. For the facial text-to-video generation task under different training conditions (*e.g.*, trained on CelebV-Text with only general appearance descriptions or

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

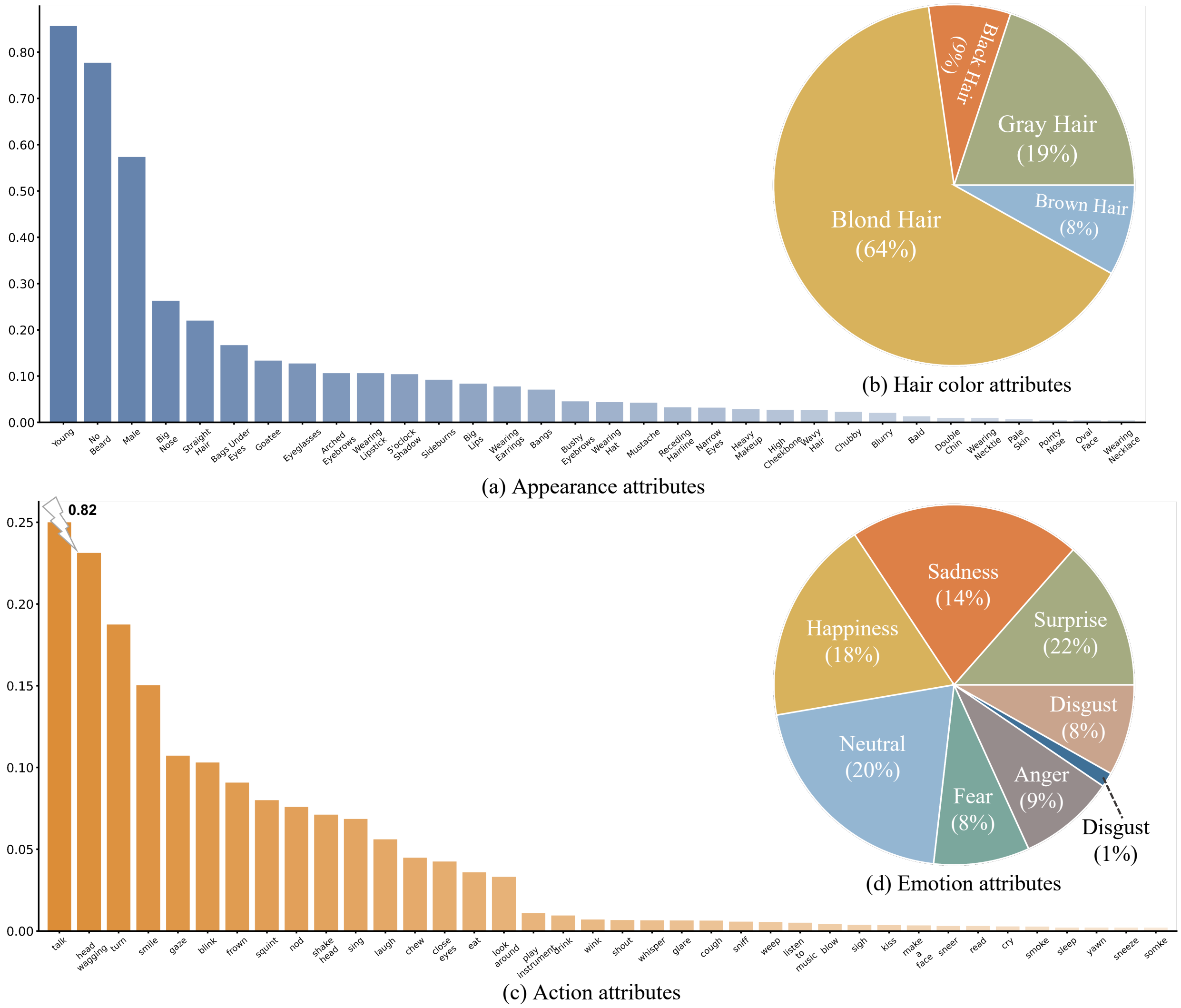


Figure A2. Distributions of general appearances, hair colors, actions, and emotions.

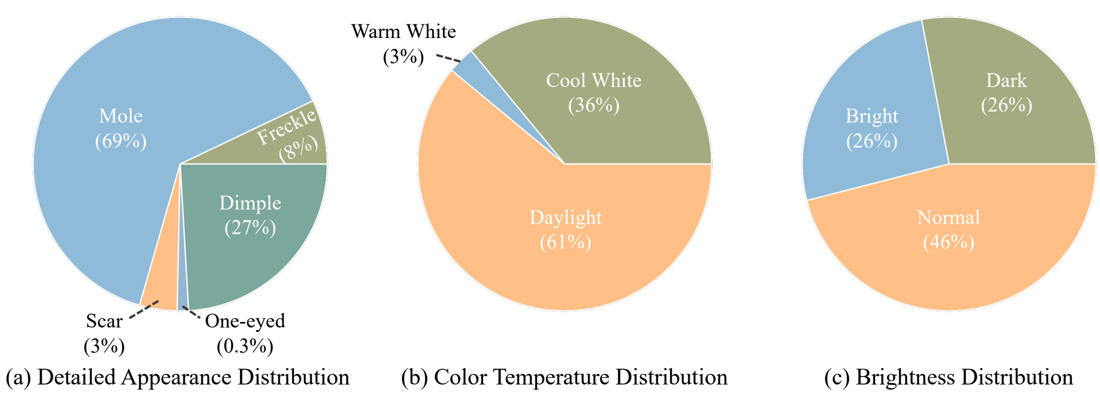
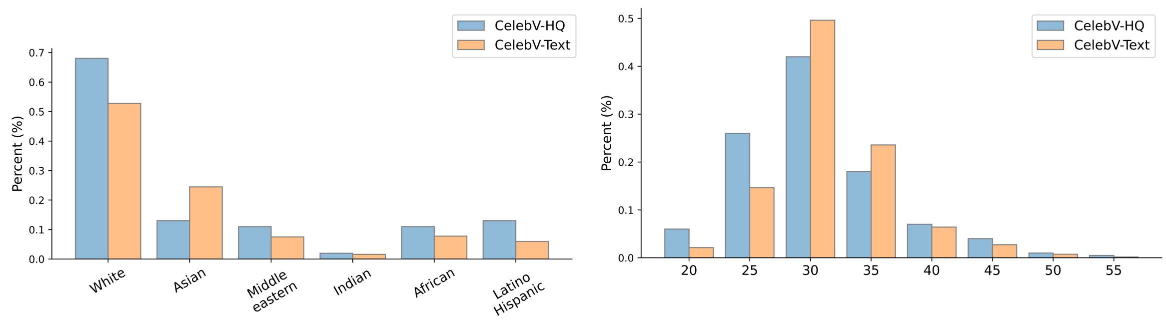


Figure A3. Distributions of detailed appearances, color temperature, and brightness.



(a) Ethnicity Distribution (b) Age Distribution

Figure A4. Distributions of ethnicity and age compared with CelebV-HQ [18].

Table A2. Detailed PCFG design for generating descriptions for general faces.

Rule		Probability
S	$\longrightarrow$ NP VP	1.0
NP	$\longrightarrow$ Det Gender	0.5
NP	$\longrightarrow$ PN	0.5
VP	$\longrightarrow$ Wearing PN Are PN HaveWith	0.166
VP	$\longrightarrow$ Wearing PN HaveWith PN Are	0.166
VP	$\longrightarrow$ Are PN HaveWith PN Wearing	0.166
VP	$\longrightarrow$ Are PN Wearing PN HaveWith	0.166
VP	$\longrightarrow$ HaveWith PN Are PN Wearing	0.166
VP	$\longrightarrow$ HaveWith PN Wearing PN Are	0.166
Wearing	$\longrightarrow$ WearVerb WearAttributes	1.0
Are	$\longrightarrow$ IsVerb IsAttributes	1.0
HaveWith	$\longrightarrow$ HaveVerb HaveAttributes	1.0
Det	$\longrightarrow$ **a**	0.333
Det	$\longrightarrow$ **the**	0.333
Det	$\longrightarrow$ **this**	0.333
Gender	$\longrightarrow$ **gender_related_attributes**	0.8
Gender	$\longrightarrow$ **person**	0.2
PN	$\longrightarrow$ **personal_noun**	1.0
WearVerb	$\longrightarrow$ **is wearing**	0.5
WearVerb	$\longrightarrow$ **wears**	0.5
WearAttributes	$\longrightarrow$ **wear_related_attributes**	1.0
IsVerb	$\longrightarrow$ **is**	1.0
IsAttributes	$\longrightarrow$ **is_related_attributes**	1.0
HaveVerb	$\longrightarrow$ **has**	0.5
HaveVerb	$\longrightarrow$ **has got**	0.5
HaveAttributes	$\longrightarrow$ **has_related_attributes**	1.0

Table A3. **Number of unique n-grams.** The numbers of unique n-grams for MM-Vox, CelebV-HQ, and CelebV-Text.

Dataset	1-grams	2-grams	3-grams	4-grams
MM-Vox [6]	65	243	1478	3935
CelebV-HQ [18]	103	372	1866	4932
CelebV-Text	**593**	**3385**	**14,136**	**45,692**

with light condition descriptions), 2,048 video samples are also generated from our trained models, which are as used as the "fake" part. To provide enough images for FID testing, 4 frames are uniformly sampled from each video. In total, we have 8192 images for the real data and fake data respectively. For both FVD and CLIPSIM evaluation, we follow [9] to generate 2048 "fake" video samples and compute the metric scores between 2048 real and fake video

Table A4. **Benchmark of text-to-video generation on different datasets.** $\downarrow$ means a lower value is better and $\uparrow$ means the opposite.

(a) Quantitative results on static descriptions, such as detailed appearance and light conditions descriptions.

Dataset	Method	FVD($\downarrow$)	FID($\downarrow$)	CLIPSIM($\uparrow$)
CelebV-Text	TFGAN [1]	415.89 $\pm$ 1.11	601.46 $\pm$ 15.12	0.155 $\pm$ 0.023
Detail App.	MMVID [6]	**68.17 $\pm$ 1.22**	**58.89 $\pm$ 5.172**	**0.191 $\pm$ 0.016**
CelebV-Text	TFGAN [1]	443.95 $\pm$ 2.23	591.00 $\pm$ 17.31	0.154 $\pm$ 0.020
Light Cond.	MMVID [6]	**69.41 $\pm$ 2.01**	**62.88 $\pm$ 4.94**	**0.187 $\pm$ 0.024**

(b) Quantitative results on dynamic descriptions of CelebV-Text.

Dataset	Method	FVD($\downarrow$)	FID($\downarrow$)	CLIPSIM($\uparrow$)
CelebV-Text **Light Dir.**	TFGAN [1]	433.02 $\pm$ 2.23	608.58 $\pm$ 16.93	0.156 $\pm$ 0.021
	MMVID [6]	69.19 $\pm$ 1.32	77.25 $\pm$ 4.05	0.172 $\pm$ 0.019
	MMVID-interp	**61.55 $\pm$ 1.28**	**60.13 $\pm$ 4.17**	**0.175 $\pm$ 0.014**
CelebV-Text **Emo.+Act.+Light Dir.**	TFGAN [1]	597.61 $\pm$ 4.96	799.14 $\pm$ 23.66	0.148 $\pm$ 0.039
	MMVID [6]	118.70 $\pm$ 3.74	107.05 $\pm$ 5.48	0.171 $\pm$ 0.023
	MMVID-interp	**100.08 $\pm$ 3.48**	**100.68 $\pm$ 5.21**	**0.173 $\pm$ 0.024**

samples. For CLIPSIM, we take the average score over all frames.

D.2. Performance Under Texts of Different Lengths

We show the model performance trained with text of different lengths while representing the same meaning in Figure A5. We discuss that lengthy inputs are closer to the distribution of the natural languages, and it is beneficial to train models with lengthy inputs due to attribute matching. Specifically, MMVID [6] trained on CelebV-Text with lengthy inputs produces satisfactory outputs when tested on short texts (Figure A5 (a)). However, outputs generated by MMVID [6] trained on MM-Vox [6] with short texts hardly reflect all attributes given long texts (e.g., straight hair in Figure A5 (b)). However, due to the limitation of baseline models, lengthy inputs would reduce the fidelity of output videos (FVD/FID in Table 4 of the main paper), which could be a new direction to devoted.

D.3. Unconditional Video Generation

To give a more comprehensive and global view of the quality of our dataset, we conduct unconditional video generation with various modern methods (*i.e.*, DIGAN [17], This young female has straight hair. She has long black hair. The woman has arched eyebrows and bags under eyes. She is first happy and then turns to be neutral. The woman smiles and then turns her head.

She is young and has straight, long black hair, arched eyebrows, bags under eyes. Happy then neutral. Smile then turn head.

(a) Qualitative results of MMVID trained on CelebV-Text with lengthy inputs. Top: long text; Bot: short text.

This young woman has straight hair. The woman has blond hair and she has got arched eyebrows. Moreover, this young woman is wearing lipstick.

This woman is young. She has straight hair, blond hair, arched eyebrows. She is wearing lipstick.

(b) Qualitative results of MMVID trained on MM-Vox with short inputs. Top: long text; Bot: short text.

Figure A5. Text-to-video generation with short and lengthy input texts.

MoCoGAN-HD [15] and StyleGAN-V [14]). Results are shown in Figure A6.

D.4. Static Face Video Generation

To further demonstrate the practical effectiveness of our CelebV-Text for facial text-video generation tasks, we additional present our generation results both quantitatively and qualitatively. As shown in Table A4 (a), results of TFGAN [1] and MMVID [6] trained on both CelebV-Text with text descriptions about detailed appearances and light conditions are listed. We can see that MMVID [6] performs better than TFGAN [1] under both conditions.

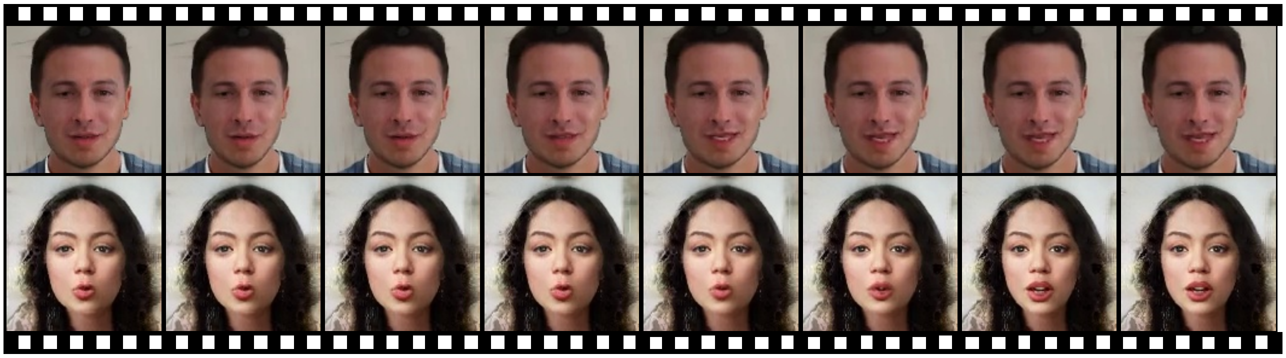
In addition, we also compare the model performance of MMVID [6] with CogVideo [10]. To validate the effectiveness of our facial text-video dataset in static attributes, we show more visualization samples in Figure A7 trained on CelebV-Text with the descriptions of static attributes (*i.e.*, detailed appearance and light conditions). We can see that although CogVideo [10] is trained on large-scale text-video dataset with larger model size than MMVID [6], MMVID [6] trained on CelebV-Text can give much better results where the generated video samples correspond well with the text input. More results by MMVID [6] trained on general appearance are shown in Figure A8. These results validate the effectiveness of our CelebV-Text.

D.5. Dynamic Face Video Generation

We show more quantitative and qualitative results when text descriptions about dynamic attributes are used for training. For all experiments, we report results of MMVID [6], MMVID-interp [6], and CogVideo [10] both quantitatively and qualitatively.

We report more quantitative results of CelebV-Text with variant input texts in Table A4 (b) and qualitative results of dynamic emotion and light direction changes in Figure A9 and Figure A10, respectively.

MMVID-interp. As mentioned in the main work, we follow [1] to apply test-time interpolation to MMVID [6] to improve text encoding and better understand the dynamics. Specifically, given the text input describing dynamic at-



(a) Qualitative results of DIGAN trained on CelebV-Text

(b) Qualitative results of MoCoGAN trained on CelebV-Text

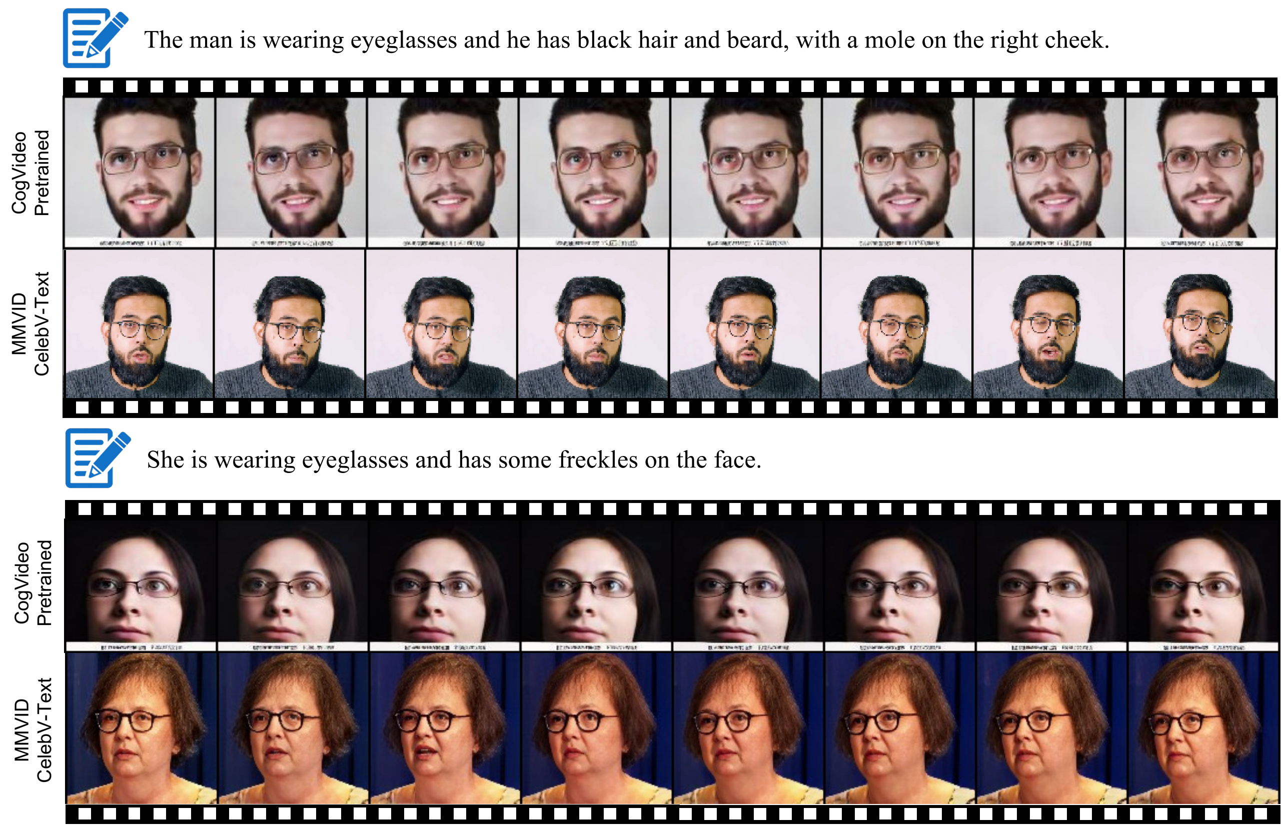
(c) Qualitative results of StyleGAN-Vtrained on CelebV-Text

Figure A6. Unconditional video generation results.

tribute changes, we manually split the dynamic description into two sentences, *i.e.*, S_1 and S_2. S_1 contains the description about the appearance and the first dynamic attribute, and S_2 contains the description about the appearance and the second dynamic attribute. Let $\mathbf{t}_{S_1}$ and $\mathbf{t}_{S_2}$ denote the feature representation obtained from the text encoder used in MMVID [11]. In this case, the description about appearance is repeated twice, so that the text encoding of it can be emphasized and improved, making the generation process more stable on preserving face identities. During the sampling process, the encoded text condition $\mathbf{t}$ is obtained by a linear interpolation between $\mathbf{t}_{S_1}$ and $\mathbf{t}_{S_2}$:

$$\mathbf{t}_i = (1-\alpha_i)\mathbf{t}_{S_1} + \alpha_i \mathbf{t}_{S_2}, \tag{1}$$

where α_i is proportional to the text sequence length. Our modification is simple and will be improved in the future.



(a) Static - Detailed Appearance

(b) Static - Light Conditions

Figure A7. **Qualitative results of facial text-to-video generation on static descriptions.** The video samples are generated given texts describing static (a) detailed appearance and (b) light conditions.

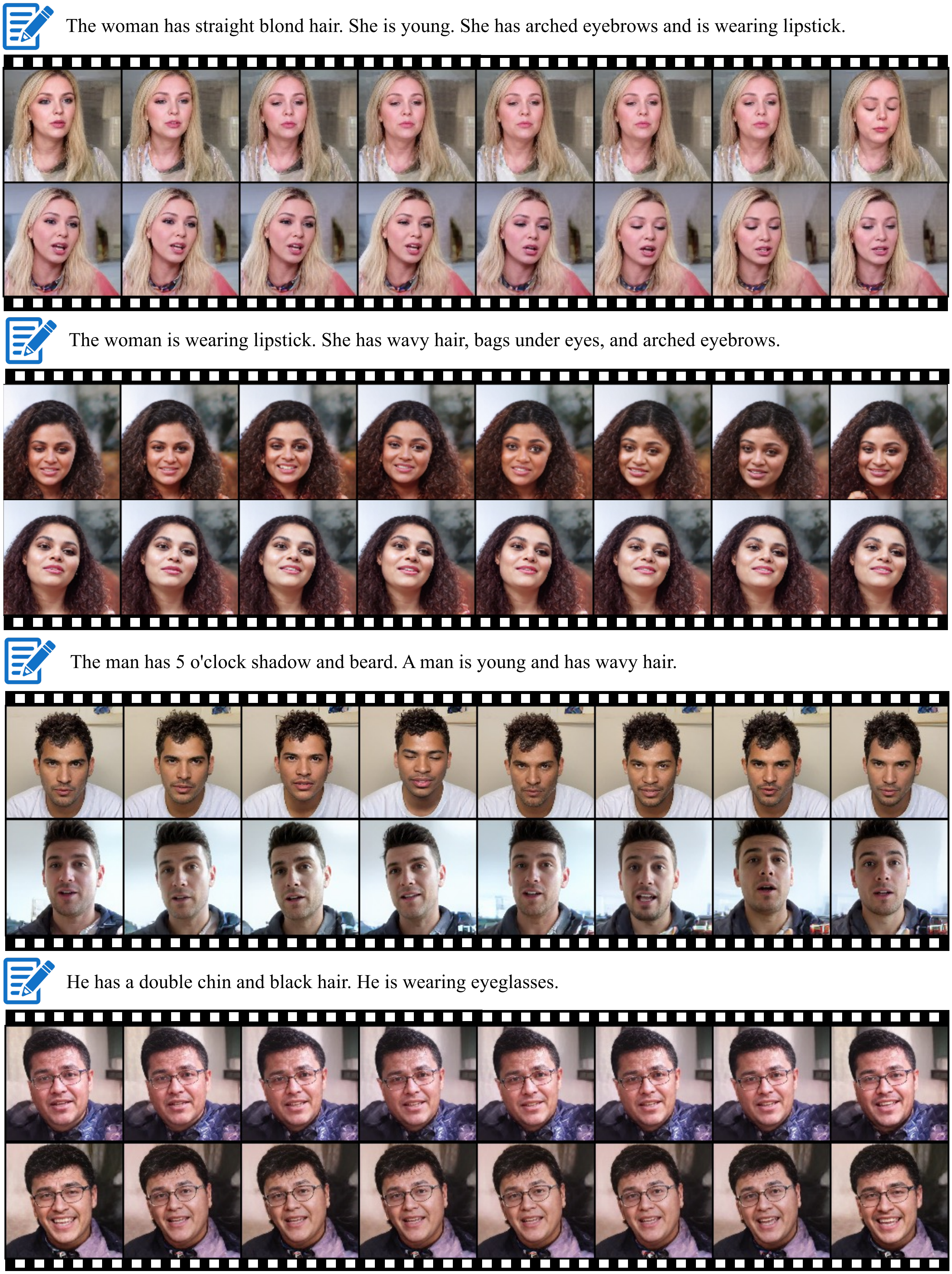


Figure A8. More sampled results from MMVID with input texts describing general appearances.

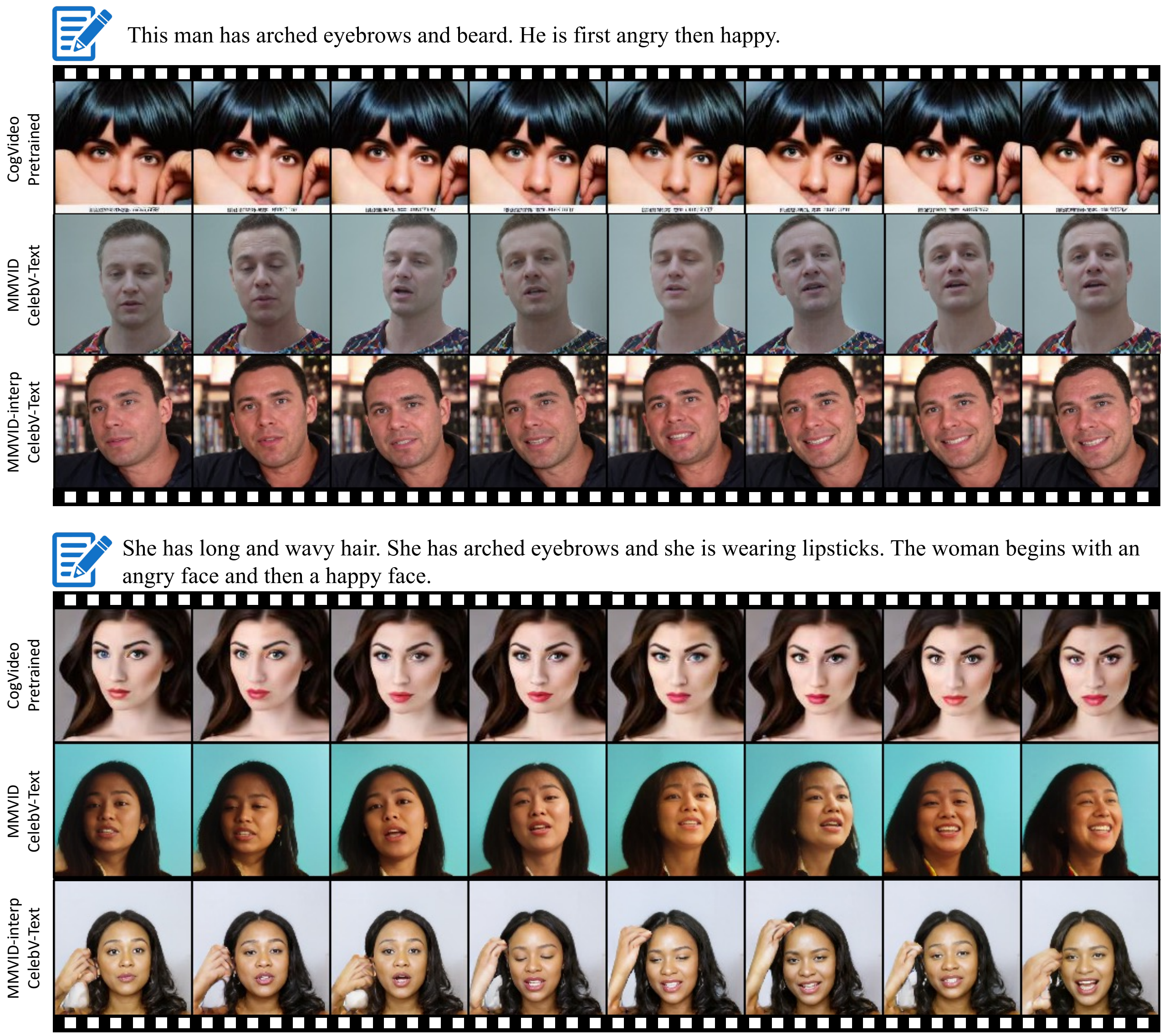


Figure A9. **Qualitative results of facial text-to-video generation.** The video samples are generated given texts describing dynamic emotion.

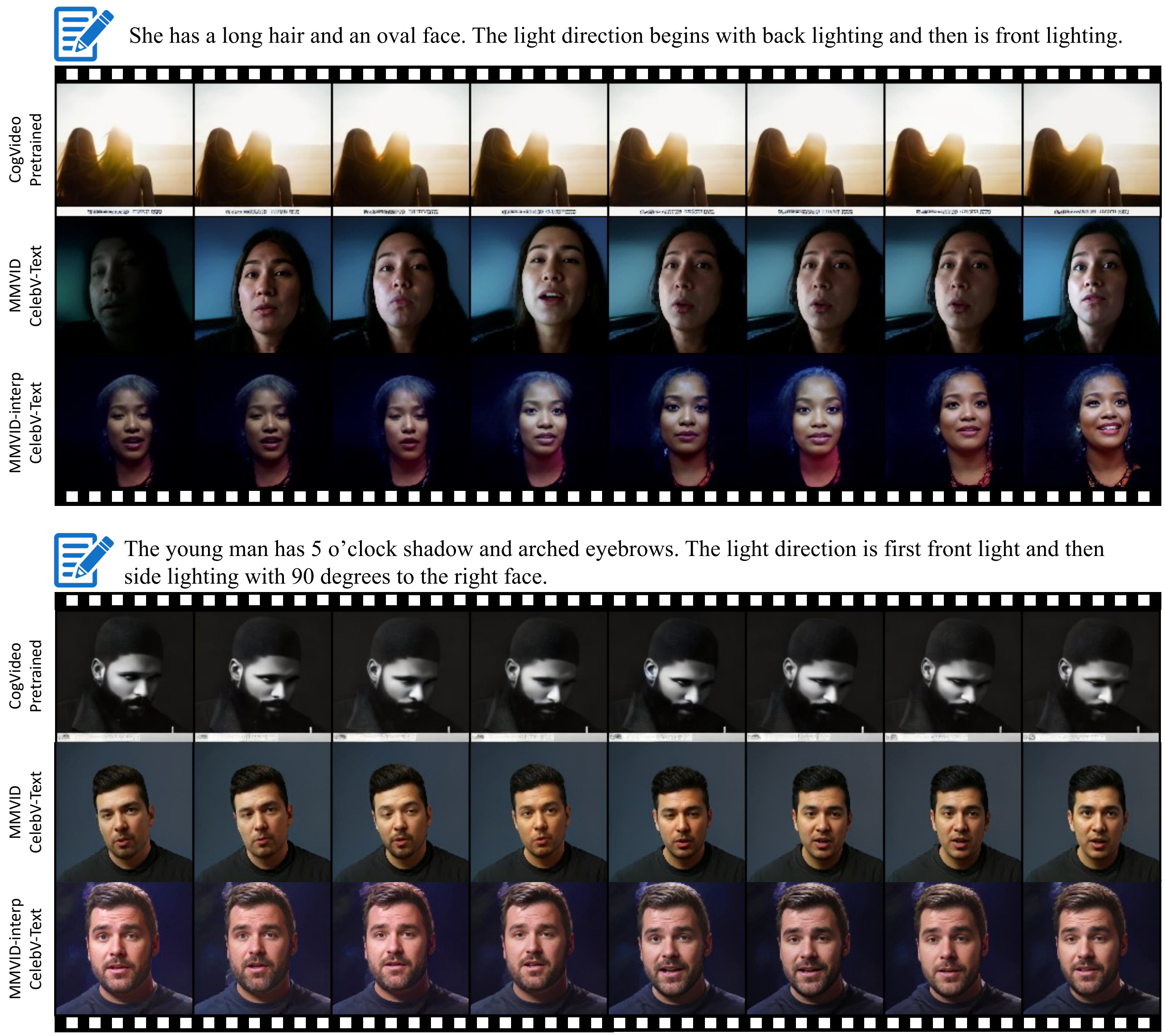


Figure A10. **Qualitative results of facial text-to-video generation on dynamic descriptions.** The video samples are generated given texts describing dynamic light directions.

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, 2019. 4, 5

[2] Sergey Bezryadin, Pavel Bourov, and Dmitry Ilinih. Brightness calculation in digital image processing. In *TDPF*, 2007. 1

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 2

[4] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, 2014. 2

[5] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *LREC*, volume 6, pages 449–454, 2006. 2

[6] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *CVPR*, 2022. 2, 4, 5

[7] Javier Hernandez-Andres, Raymond L Lee, and Javier Romero. Calculating correlated color temperatures across the entire gamut of daylight and skylight chromaticities. In *Applied optics*, 1999. 1

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 2

[9] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4

[10] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 5

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[12] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022. 1

[13] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 1964. 1

[14] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 5

[15] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021. 5

[16] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[17] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. 4

[18] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 1, 2, 4