DH-FaceVid-1K: A Large-Scale High-Quality Dataset for Face Video Generation

Supplementary Material

Dataset Construction

A. Data details.

The majority of the facial video data in this dataset, over 65%, consists of single-person frontal interview videos, ensuring the quality and stability of the raw data. To enhance the diversity of video scenes, we also collected some outdoor vlog-style talking head video data. To ensure the diversity of the dataset, in over 95% of the total data, each identity corresponds to no more than 10 video clips.

As mentioned in the manuscript, we applied the data processing pipeline described in the text to filter the public datasets CelebV-HQ, CelebV-Text, and TalkingHead-1KH. The filtered sample data can be found in Figure 2.

B. CodeFormer bias

CodeFormer-enhanced videos sometimes exhibit artifacts, particularly around the eyes of Asian faces. To address this, a team of 10 manually **reviewed** the first frame of each video, discarding any that display such issues. The demo data corresponding to these failures can be found in Figure 1.

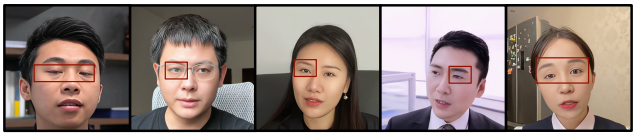

Figure 1. Sample image showing artifacts in the eye area after introducing CodeFormer.

Figure 2. Filtered out sample data from public **datasets**.

C. Video Annotations

We first **used** PLLaVA to automatically generate video annotations. More specifically, we **interacted** with PLLaVA in a manner similar to conversing with an LLM (Large Language Model). We **provided** it with a template prompt,

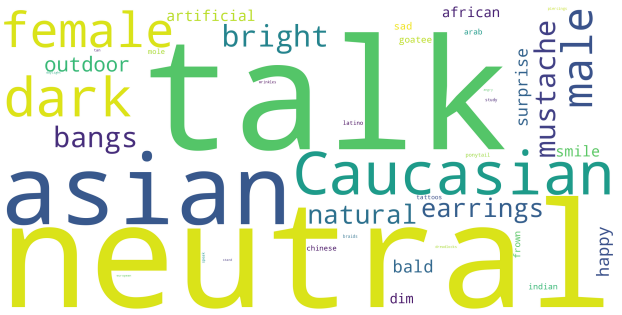

Figure 3. **Keywords** of DH-FaceVid-1K's annotations.

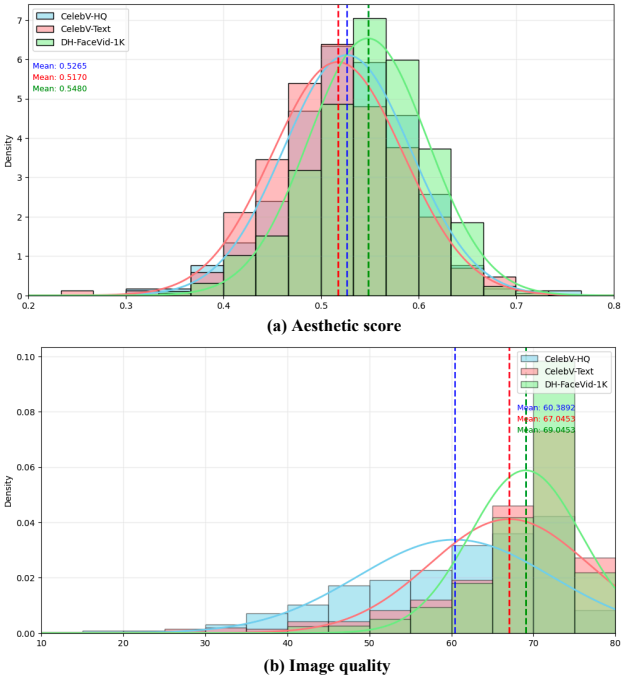

Figure 4. Comparison of CelebV-HQ, CelebV-Text**,** and **DH-FaceVid-1K** on VBench.

which it then **used** to generate corresponding video annotations. Our annotations primarily include information about appearance, expressions, actions, and the environment—details that are also embedded within the template, where we **provided** options for various attributes of these annotations. For example, for the attribute **"emotion"**, we **required** selecting from options such as Neutral, Happy, Sad, and Angry. The selection of these options is based on the attribute design from CelebV-Text.

Table 1. Comprehensive attribute list of DH-FaceVid-1K, including ethnicities, appearance details, emotions, actions, and lighting conditions.

Static Attributes					
(a) Ethnicities					
Asian	White	Indian	European	African	Arab
Latino	Chinese	Japanese	Korean	Latina	Jewish
Mexican					
(b) General Appearance					
No beard	Male	Female	Straight hair	Black hair	Anchored eyebrow
Short hair	Eyeglasses	Brown hair	Rosy cheeks	Long hair	Bush eyebrow
Chubby	Heavy makeup	Oval face	Lipsticks	Big nose	Wavy hair
Bangs	Double chin	Side burn	Earrings	Goatee	White hair
Blonde hair	Bald	Old	Narrow eyes	Big lips	Pale skin
Wearing shirt	Wearing jacket	Wearing suit	Wearing tie	Wearing coat	Wearing blazer
Wearing hoodie	Redhead	Wearing cloak	Wearing bonnet	Blemish	Pompadour
Mustache	Wearing necklace	Wearing sweater	Wearing blouse	Wearing cap	Wearing scarf
Wearing hat	Wearing dress	Wearing headband	Mole	Wearing vest	Tattoos
Wearing sunglasses	Piercings	Wearing beanie	Wearing hijab	Ponytail	Scars
Wearing robe	Tan	Wearing ring	Wrinkles	Wearing hood	Wearing backpack
Wearing bandana	Braids	Wearing turban	Pigtails	Dreadlocks	With purse
(c) Light Conditions					
Dark	Outdoor	Bright	Natural	Artificial	Dim
Daylight	Normal				
Dynamic Attributes					
(a) Action					
Talk	Smile	Wag Head	Frown	Gaze	Nod
Shake Head	Sing	Glare	Close Eyes	Laugh	Whisper
Wink	Shout	Sigh	Listen To Music	Yawn	Cough
Sneeze	Study	Speak	Sleep	Kiss	Pile
Stand	Think	Eat	Drink	Read	Touch
Roll	Rest	Work	Wash	Ride	Drive
Play	Hold	Throw	Run	Clutch	Tilt
Hug	Walk				
(b) Emotion					
Neutral	Happy	Sad	Angry	Fear	Surprise
Contempt	Disgust				

After obtaining the initial annotations, we **employed** an additional 50 annotators to review the text. These annotators **were tasked** with assessing static information in the annotations, such as age, ethnicity, clothing, and facial accessories, based on the first frame of the video. Subsequently, they **evaluated** dynamic information, such as emotional and movement changes, based on three randomly selected non-consecutive additional frames.

The comprehensive attribute list of DH-FaceVid-1K is presented in Table 1**.** We visualized some high-frequency words in the annotations. The word cloud can be found in Figure 3.

D. Challenges

Although our primary goal is to construct a high-quality talking face dataset mainly composed of Asians to address the scarcity of such datasets, the statistical data presented in the main text **indicated** a certain long-tail distribution in racial representation. However, through experiments, we discovered that by integrating our proposed dataset with cleaned open-source data and thoroughly training powerful video generation models, such as CogVideoX, it **was** still easy to generate high-quality data for Caucasian, African, and Indian individuals.

In terms of actions and emotions, talking and neutral emotions **dominated** overwhelmingly. Nevertheless, we **found** that even a small proportion of data **was** sufficient for the model to learn, given our total duration of nearly 1200 hours. Qualitative experimental images for other ethnicities, scenes, emotions, and actions, such as smiling, can be found in the later sections.

E. Training Protocol

To validate the effectiveness and superiority of the proposed dataset, we evaluated the performance of generative models under different architectures based on T2V (Text-to-Video) and **I2V (Image-to-Video)** in the earlier experimental sections. All model training **was conducted** separately using CelebV-Text and DH-FaceVid-1K as the training datasets.

We **aimed** to replicate these methods according to the training details provided officially, including the optimizer, learning rate, and loading the official weight checkpoints. All models **were fine-tuned** for 100k steps using eight A800 GPUs. However, for SVD, since the official training code is not open-sourced, we **trained** it based on a third-party reproduced code.

For OpenSora, we **fine-tuned** using version 1.2. For EasyAnimate, we **used** version v3 for fine-tuning. However, this version often **encountered** crashes when fine-tuning with large datasets. Our solution **was** to reduce the learning rate (down to 2e-6) and frequently save checkpoints, allowing us to revert to the most recent stable checkpoint in case of a crash.

F. Quantitative Experiment Settings

We **utilized** FVD, FID, and **CLIPScore** to evaluate video temporal consistency, individual frame quality, and the relevance between the generated video and text prompt. Before training different models, we randomly **selected** 2048 videos from each of the two training datasets as the test set, which **were** not involved in training. For Text-to-Video (T2V) generation experiments, we **generated** 2048 videos as the "fake" videos to calculate FVD and **CLIPScore**, based on the prompts of the test set. Subsequently, to compute FID, we uniformly **sampled** 4 frames from each of these 2048 "fake" videos as fake data and **applied** the same procedure to the test data. In total, we **had** 8192 images for both real and fake data. For Image-to-Video (I2V), we **took** the first frame from the test set images as the image prompt to **generate** 2048 videos as fake data, with other settings similar to the T2V experiments.

To ensure the fairness of the comparison, all real and fake data **were** uniformly resized to 512×512.

G. More Qualitative Results

We present additional qualitative results that were not covered in the main text. For the Text-to-Video experiments, we first **used** more diverse prompts, such as different emotions, ethnicities, actions, and scenes, to generate results. We then **included** generation results from prompts with formats different from those in the dataset annotations, such as prompts lacking key information or using single-word descriptions of character traits in the video. This demonstrates that our model, after thorough training, possesses strong generalization capabilities. For Image-to-Video (I2V), we **used** some **"wild"** data to generate results, verifying that the model, trained on our dataset, still exhibits good extrapolation abilities. The additional qualitative experimental results in this section **were generated** using the CogVideoX 5b version and **OpenSora**.

H. Comparison using Vbench

We conducted a comparative analysis of aesthetic scores and image quality metrics between DH-FaceVid-1K, CelebV-HQ, and CelebV-Text using VBench [1], as illustrated in Figure 4. The results demonstrate that the proposed DH-FaceVid-1K dataset achieves comparable aesthetic performance while exhibiting superior image quality relative to the benchmark datasets.

Table 2. Comparison of T2V metrics on cross-dataset evaluation.

Dataset	FID (↓)	FVD (↓)	CLIP (↑)	DSL-FIQA (↑)
CelebV-Text	25.71	313.51	0.838	0.7531
Ours	**22.81**	**250.53**	**0.845**	**0.8204**

Table 3. User study measured by Mean Opinion Scores.

Datasets	Adherence	Naturalness	Visual Quality	Overall
CelebV-HQ	2.53	4.01	3.75	3.90
CelebV-Text	2.89	3.88	4.17	4.01
Ours	**4.52**	**4.39**	**4.33**	**4.47**

I. Combination with Existing Datasets

Our dataset includes 200h of Caucasian and African data, with diverse generation examples shown in Appendix Figures 13-16. Given that existing public datasets already contain extensive non-Asian data (about 4.5k h), we recommend using either equal-duration sampling (1:1) or **population-proportional splits** to ensure generalization and racial diversity.

J. Cross Reference-Dataset Evaluation

We **conducted** cross-dataset evaluation using CogVideoX-2B T2V to assess the generalization capability of our dataset. The evaluation sets include HDTF and CelebV-HQ **(at a 1:3 ratio, with results reported in Table 2)**. Our dataset demonstrates strong generalizability and consistently outperforms CelebV-Text across all metrics.

K. User Study

We **conducted** a user study with 25 participants to evaluate CogVideoX's T2V face generation performance after fine-tuning. Participants **compared** 30 videos generated with the same prompt by models trained on CelebV-HQ, CelebV-Text, and the proposed dataset. The prompts **included** diverse ethnic groups and **emotions (neutral:happy:angry:sad = 6:1:1:1)** in various settings. Evaluations **focused** on prompt adherence, face naturalness, video quality, and overall quality. As shown in Table 3, videos using our fine-tuned weights **scored** the highest overall.

L. Future Work

Our proposed dataset primarily focuses on talking faces, and when combined with powerful open-source models, it already enables robust and diverse talking face generation capabilities. However, there are still limitations, such as the video's restriction to facial expressions without including common actions like eating or crying. On the other hand, our current dataset contains a limited number of non-East

Asian faces. To further enhance data diversity, we are actively collecting additional Asian facial data from regions beyond East Asia (including Indian, Indonesian, Malay, and Vietnamese populations), while strictly adhering to data privacy protocols.

Our future work will focus on developing datasets that support digital human generation for both half-body and full-body models, aiming to achieve true digital human synthesis.

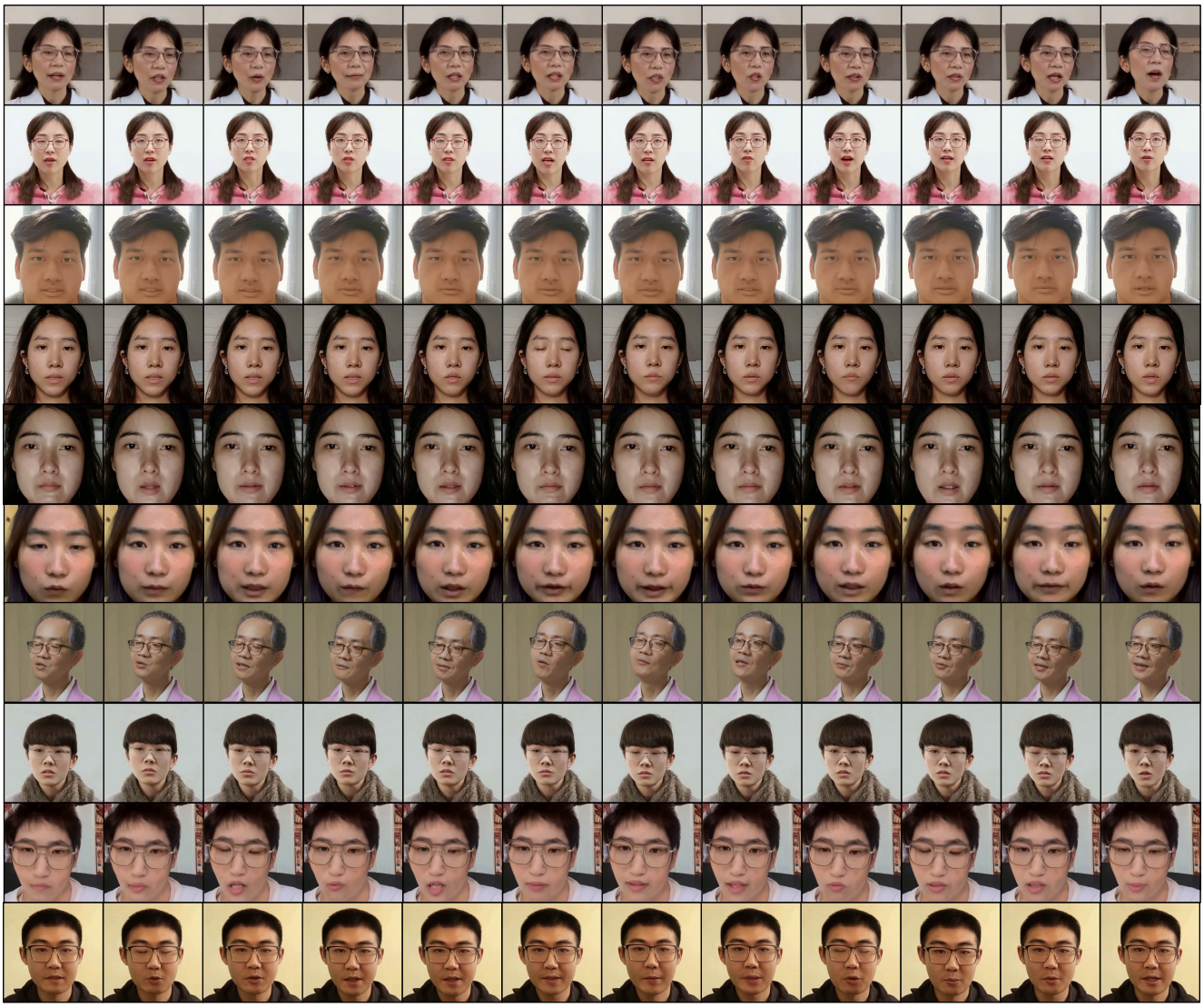

Figure 5. More qualitative T2V evaluation results from CogVideoX demonstrate the effectiveness of the DH-FaceVid-1K fine-tuned model in **generating videos** from different angles and facial distances.

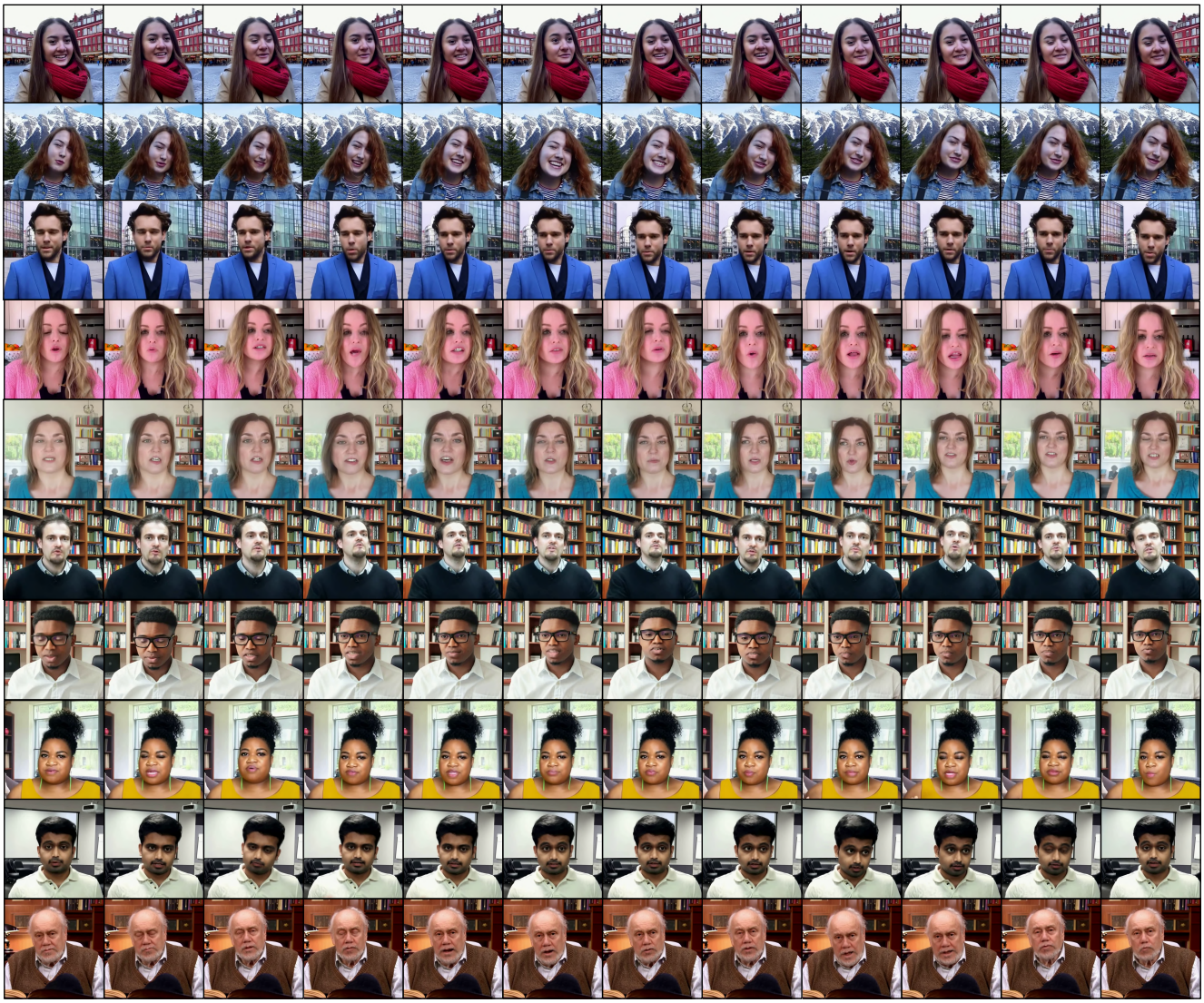

Figure 6. More qualitative T2V evaluation results from OpenSora confirm that the DH-FaceVid-1K fine-tuned model retains the ability to generate speaking videos of faces from multiple ethnicities, including Caucasian, African, and Indian.

Figure 7. More qualitative I2V evaluation results from CogVideoX confirm that the DH-FaceVid-1K fine-tuned model retains the ability to generate various styles (realistic and cartoon) across multiple ethnicities.

References

[1] Ziqi Huang, Yinan He, Jiashuo Yu, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 3

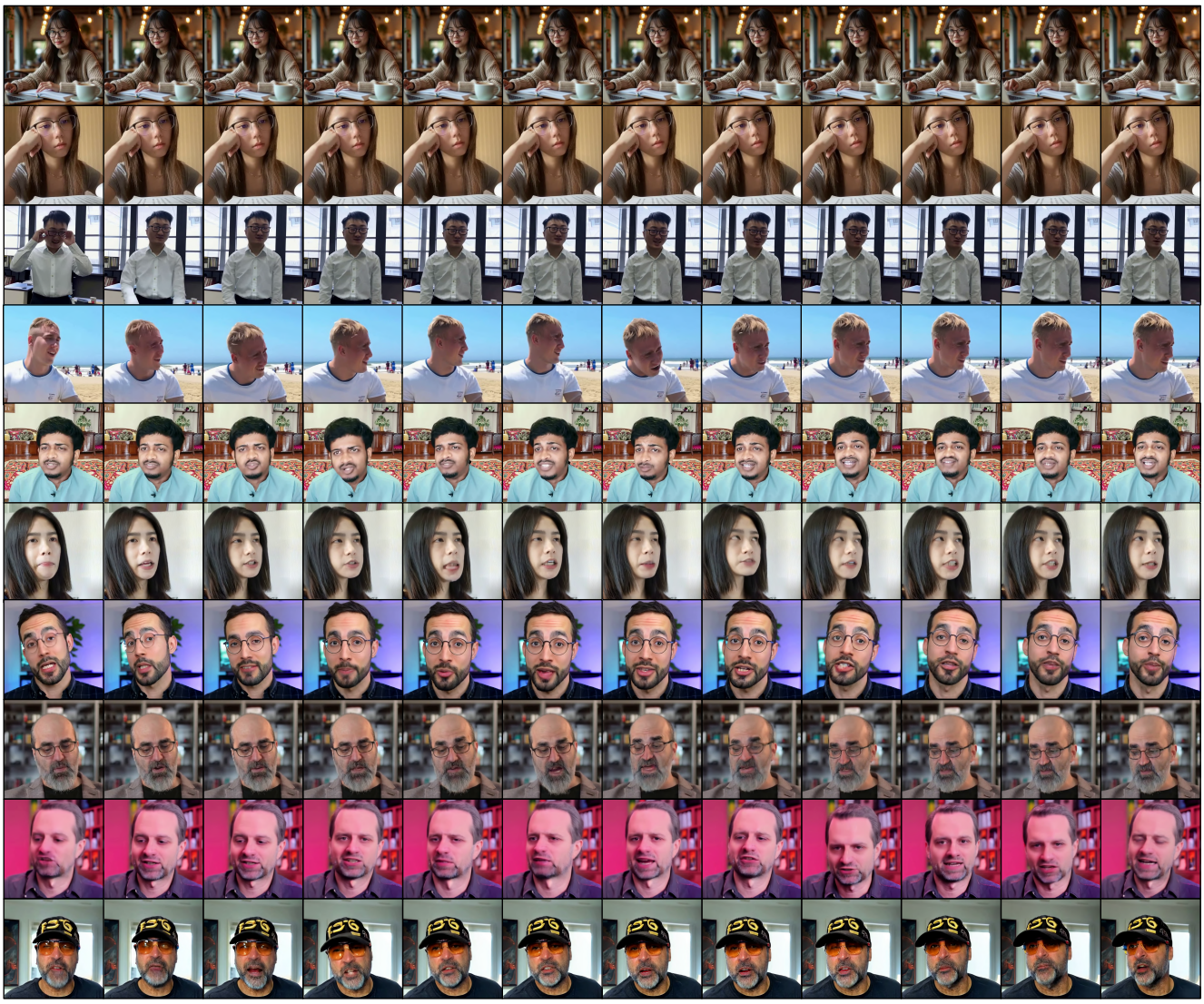

Figure 8. More qualitative I2V evaluation results from CogVideoX confirm that the DH-FaceVid-1K fine-tuned model retains the ability to generate not only speaking but also various other actions, such as reading, and can produce half-body images beyond just the face, across multiple ethnicities.