CelebV-HQ: A Large-Scale Video Facial Attributes Dataset

Hao Zhu[1*], Wayne Wu[1*✉], Wentao Zhu[2], Liming Jiang[3], Siwei Tang[1], Li Zhang[1], Ziwei Liu[3], and Chen Change Loy[3]

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

(a) Appearance (b) Action (c) Emotion

Fig. 1: **Overview of CelebV-HQ.** CelebV-HQ contains 35,666 videos, including 15,653 identities. Each video was manually labeled with 83 facial attributes, covering appearance, action, and emotion attributes.

Abstract. Large-scale datasets have played indispensable roles in the recent success of face generation/editing and significantly facilitated the advances of emerging research fields. However, the academic community still lacks a video dataset with diverse facial attribute annotations, which is crucial for the research on face-related videos. In this work, we propose a large-scale, high-quality, and diverse video dataset with rich facial attribute annotations, named the High-Quality Celebrity Video Dataset (CelebV-HQ). CelebV-HQ contains 35,666 video clips with the

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

resolution of 512×512 at least, involving $15,653$ identities. All clips are labeled manually with 83 facial attributes, covering appearance, action, and emotion. We conduct a comprehensive analysis in terms of age, ethnicity, brightness stability, motion smoothness, head pose diversity, and data quality to demonstrate the diversity and temporal coherence of CelebV-HQ. Besides, its versatility and potential are validated on two representative tasks, *i.e.*, unconditional video generation and video facial attribute editing. We finally envision the future potential of CelebV-HQ, as well as the new opportunities and challenges it would bring to related research directions. Data, code, and models are publicly available[4].

Keywords: Large-scale video dataset; Facial attribute annotation; Face video generation and editing

1 Introduction

The rapid development of Generative Adversarial Networks (GANs) [17,55,34,36,37,35] has demonstrably promoted advances in face generation and editing. This progress relies heavily on the contribution of large-scale datasets, *e.g.*, CelebA [45], CelebA-HQ [34], and FFHQ [36]. These datasets, with high-quality facial images, have facilitated the development of a series of face generation and editing tasks, such as unconditional face generation [17,55,36,37,65,81,63,31], facial attribute editing [8,26,61,66,77] and neural rendering [24,18,52,4,52,3,19,14,6]. However, most of these efforts are based on static *image modality*. In industry, with the booming development of mobile internet [10], *video modality* data begins to take a bigger and bigger share in customers' daily shootings [28,48]. A well-suited dataset, which is capable of supporting the face generation and editing tasks in video modality, is eagerly asked.

Recent works [37,2] have shown that the *scale* and *quality* are essential factors for a facial dataset in image modality. A more sufficient utilization of large-scale datasets would improve model generalization [58], while the quality of the dataset largely determines the limit of the generative models [55,36,37,65,81,63]. In addition, facial *attribute* provides effective information to help researchers go more deeply into the face-related topics [45,26,8,61]. However, the current public facial datasets consist of either static images with attribute labels [45,34] or videos with insufficient scale [72] and quality [51,9].

Constructing a large-scale and high-quality face video dataset with diverse facial attribute's annotations is still an open question, given the challenges brought by the nature of video data. 1) Scale. The collected videos need to meet several requirements, such as temporal consistency, high-resolution and full-head. The strict standards together with the limited sources, make the expansion of dataset's scale both time and labor consuming. 2) Quality. The quality is not

[4] Project page: `https://celebv-hq.github.io/`
Code and models: `https://github.com/CelebV-HQ/CelebV-HQ`

only reflected in the high fidelity and resolution, but also in the diverse and natural distribution of data samples. It asks for a well-designed data-filtering process to ensure all of the requirements of fidelity, resolution and data distribution. 3) Attribute Annotation. The coverage of the facial attribute set need to be sufficient to describe a human face thoroughly, both in the time-invariant and time-variant perspective. Also, the annotation process need to be accurate and highly efficient.

In order to tackle the challenges discussed above, we carefully devise a procedure for dataset construction. First, to ensure the scale of the collected video, we build a large and diverse set of Internet queries. The designed queries cover a rich set of scenarios and thus successfully enable a huge raw data pool with millions of clips. Then, to filter out high-quality data from the raw data pool, we introduce an automatic pre-processing pipeline. In this pipeline, we leverage face detection and alignment tools to ensure the high fidelity and resolution. Finally, we propose a facial attributes set with extensive coverage, including appearance, action and emotion. To ensure the accuracy and efficiency of the annotation, we design a systematic attributes annotation process, including annotator training, automatic judgment and quality check steps.

To this end, we successfully create the High-Quality Celebrity Video (CelebV-HQ) Dataset, a large-scale, diverse, and high-quality video facial dataset with abundant attributes' annotations. CelebV-HQ contains $35,666$ in-the-wild video clips with the resolution of 512×512 at least, involving $15,653$ person identities and 83 manually labeled facial attributes. Our labeling comprises a comprehensive set of face-related attributes, including 40 appearance attributes, 35 action attributes, and 8 emotion attributes. Samples on CelebV-HQ are shown in Fig. 1.

We perform a comprehensive analysis of data distribution to demonstrate CelebV-HQ's statistical superiority to existing image and video datasets. First, compared to image datasets with attribute annotations [45,34], CelebV-HQ has much higher resolution ($2\times$) than CelebA [45] and comparable scale to high-quality dataset [34]. Also, by comparing CelebV-HQ with CelebA-HQ [34] in the *time-invariant* aspects, we demonstrate that CelebV-HQ has a reasonable distribution on appearance and facial geometry. Furthermore, we compare CelebV-HQ with a representative video face dataset VoxCeleb2 [9] in the *time-variant* aspects, such as temporal data quality, brightness variation, head pose distribution, and motion smoothness, suggesting that CelebV-HQ has superior video quality.

Besides, to demonstrate the effectiveness and potential of CelebV-HQ, we evaluate representative baselines in two typical tasks: unconditional video generation and video facial attribute editing. For the task of unconditional video generation, we train state-of-the-art unconditional video GANs [65,81] on CelebV-HQ fullset and its subsets that divided by different actions. When trained on different subsets of CelebV-HQ, the corresponding actions can be successfully generated. Further, we explore the video facial attribute editing task using temporal constrained image-to-image baselines [26,8]. Thanks to the rich sequential information included in CelebV-HQ dataset, We show that simple modification of current image-based methods can bring remarkable improvement in the tem-

Table 1: **Face datasets comparison.** The symbol "#" indicates the number. The abbreviations "Id.", "Reso.", "Dura.", "App.", "Act.", "Emo.", "Env.", and "Fmt." stand for Identity, Resolution, Duration, Appearance, Action, Emotion, Environment, and Format, respectively. The "∗" denotes the estimated resolution.

Datasets	**Meta Infomation**				**Attributes**			**Env.**	**Fmt.**
	#Samples	#Id.	Reso.	Dura.	App.	Act.	Emo.		
CelebA [45]	202,599	10,177	178×218	N/A	✓	✗	✗	Wild	IMG
CelebA-HQ [34]	30,000	6,217	1024×1024	N/A	✓	✗	✗	Wild	IMG
FFHQ [36]	70,000	N/A	1024×1024	N/A	✗	✗	✗	Wild	IMG
CelebV [76]	5	5	256×256	2hrs	✗	✗	✗	Wild	VID
FaceForensics [56]	1,004	1,004	256×256∗	4hrs	✗	✗	✗	Wild	VID
VoxCeleb [51]	21,245	1,251	224×224	352hrs	✗	✗	✗	Wild	VID
VoxCeleb2 [9]	150,480	6,112	224×224	2,442hrs	✗	✗	✗	Wild	VID
MEAD [72]	281,400	60	1980×1080	39hrs	✗	✗	✓	Lab	VID
CelebV-HQ	35,666	15,653	512×512	68hrs	✓	✓	✓	Wild	VID

poral consistency of generated videos. The experiments conducted above empirically demonstrate the effectiveness of our proposed CelebV-HQ dataset. Additionally, CelebV-HQ could potentially benefit the academic community in many other fields. We provide several empirical insights during constructing CelebV-HQ dataset and make an exhaustive discussion of the potential of CelebV-HQ in research community.

In summary, our contributions are threefold: 1) We contribute the first large-scale face video dataset, named CelebV-HQ, with high-quality video data and diverse manually annotated attributes. Corresponding to CelebA-HQ [34], CelebV-HQ fills in the blank on video modality and facilitates future research. 2) We perform a comprehensive statistical analysis in terms of attributes diversity and temporal statistics to show the superiority of CelebV-HQ. 3) We conduct extensive experiments on typical video generation/editing tasks, demonstrating the effectiveness and potential of CelebV-HQ.

2 Related Work

2.1 Video Face Generation and Editing

Recent advances in face video generation typically focused on unconditional video generation [70,57,67,81,65,63] and conditional face video generation [76,62,82,73,85,5,88,86,30]. Conventional unconditional video face generation [70,57,67,81] are mainly based on GANs [17]. These models usually decompose the latent code into content and motion codes to control the corresponding signals. Some recent efforts [65,63] aimed to extend high-quality pre-trained image generators to a video version to exploit the rich prior information. Conditional face video generation mainly including face reenactment [76,62,82,73] and talking face generation [85,5,88,86,30]. The motivation of these tasks is to use visual and audio modalities to guide the motion of a face video.

Face video editing is another emerging field [79,68]. The common characteristic of these works is to edit face attributes on the StyleGAN [37] latent space. Nevertheless, due to the lack of large-scale high-quality video datasets, these video-based editing efforts are still trained on images, exploiting the rich information of a pre-trained image model [37]. This leads to the main problem of having to solve for temporal consistency. The face video dataset proposed in this paper would help to address such hurdles and facilitate more interesting research in video face generation and editing.

2.2 Face Datasets

Face datasets can be divided into two categories: image datasets and video datasets. Many face image datasets are initially proposed for face recognition, like LFW [25] and CelebFaces [64] which largely promote the development of related fields. To analyze facial attributes, datasets like CelebA and LFWA [45] have been proposed. Both of them have 40 facial attribute annotations and have advanced the research field to a finer level of granularity. CelebA-HQ [34] improves 30k images in CelebA to 1024×1024 resolution. CelebAMask-HQ [38] further labels 19 classes of segmentation masks. CelebA-Dialog [32] labels captions describing the attributes.

In addition to the above image datasets, many video datasets have also been released. CelebV [76] was proposed for face reenactment. Audiovisual datasets such as VoxCeleb [51] and VoxCeleb2 [9] were originally released for speaker recognition, and further stimulated the development of audiovisual speaker separation and talking face generation domains. There are also several face video datasets with emotion attributes, such as RAVDESS [46] and MEAD [72]. MEAD [72] is the largest emotional video dataset, which includes 60 actors recorded from seven view directions. However, all of these datasets either contains only images with attribute annotations or are unlabeled videos with insufficient diversity. The rapidly growing demand for video facial attribute editing cannot be met. A video version of the dataset like CelebA-HQ [34] is urgently needed.

3 CelebV-HQ Construction

A dataset lies the foundation for model training, and its quality greatly affects the downstream tasks. The principle of building CelebV-HQ is to reflect real-world distribution with large-scale, high-quality, and diverse video clips. Hence, we design a rigorous and efficient pipeline to construct CelebV-HQ dataset, including Data Collection, Data Pre-processing, and Data Annotation.

3.1 Data Collection

The data collection process consists of the following steps. We start by creating various queries in order to retrieve human videos that are diverse in content

and rich in attributes. The queries are designed to include keywords of different categories such as celebrity names, movie trailers, street interviews and vlogs, all in different languages, with 8376 entities, and 3717 actions. Then, we use these queries to collect the raw videos from the Internet. During the collection process, we introduce several constraints to discard unsatisfactory videos. For each query, we only collect the first 30 results to reduce duplicate human IDs, and the raw videos are required to have a resolution greater than 1080p with a normal bitrate. Consequently, we obtain a raw data pool with millions of video clips.

3.2 Data Pre-processing

In order to sample high-quality face video clips from the raw data pool, we develop an automatic video pre-processing pipeline. Please refer to the supplementary materials for more details. We choose 512^2 as the normalized resolution due to the following reasons. 1) The face regions of web videos usually do not reach the resolution of 1024^2 or higher, and it is difficult to obtain super high-resolution videos and ensure their diversity. Before the rescaling, the percentage of video resolution: 0.6% for $450^2 \sim 512^2$, 76.6% for $512^2 \sim 1024^2$, and 22.7% for 1024^2+. 2) We need to make sure that all the videos are of the same resolution when training models. We choose 512^2 to ensure that all the clips are not upsampled significantly, which would affect the video quality. Also, to meet different usage scenario, a tool is provided on our project page that offers options to keep the original resolution.

3.3 Data Annotation

Data annotation is the core part of CelebV-HQ, and the annotation accuracy is vital. We first describe how we select the attributes to be annotated, then present the standard protocol of manual annotation.

Attribute Selection. We decouple a face video into three factors, *i.e.*, appearance, action, and emotion. Appearance describes the facial attributes that do not change along with the video sequence, such as hair color and gender. Action describes facial attributes that are related with video sequence, such as laugh and talk. Emotion describes the high-level mental status of human, such as neutral and happy. These three categories serve as important feature dimensions to characterize face video clips. We provide the design details and the complete list of all the attributes in the supplementary materials.

Attribute Annotation. To ensure the accuracy of the annotations, our entire annotation process includes the training of annotators, annotation, and quality control. Before the labeling begins, training courses are provided to help annotators understand each attribute and to have the same criteria for judging each attribute. We set up a Multi-label Annotation Table for each video, the table contains all labels that need to be labeled. Each video clip is independently annotated by 5 trained annotators. We select the annotation that has been agreed the most. If the annotation is only marginally agreed (3 vs 2), the sample will

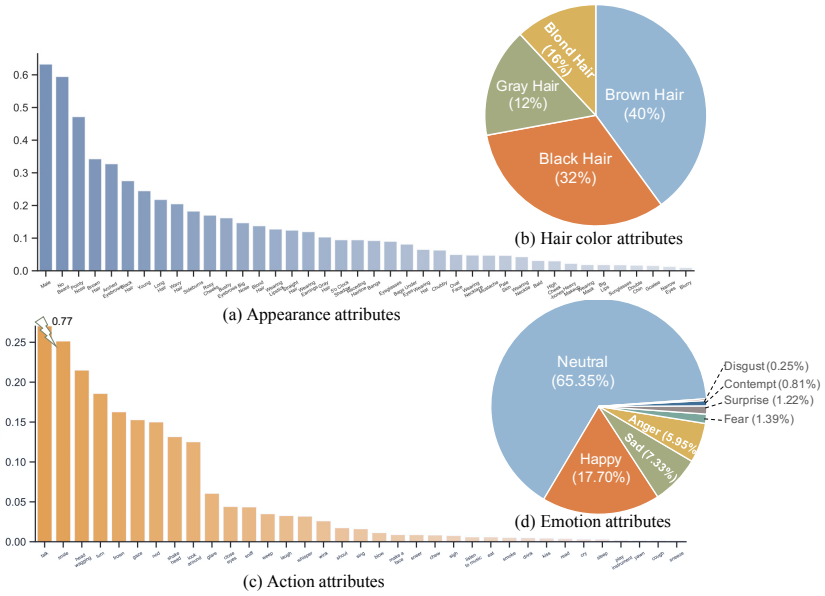


Fig. 2: **The distributions of each attribute.** CelebV-HQ has a diverse distribution on each attribute category. (Please zoom in for details).

be re-labeled. Finally, we additionally take a Quality Check process, in which the annotated data is further inspected by a professional quality inspector. If the annotated data does not meet the standard, it will also be re-labeled.

4 Statistics

In this section, we present the statistics of CelebV-HQ to demonstrate its statistical superiority. Then, we make comparison of CelebV-HQ with two most related and representative image and video datasets (*i.e.*, CelebA-HQ [34] and VoxCeleb2 [9]) respectively, in which we verify that the proposed CelebV-HQ has a natural distribution and better quality.

4.1 Analysis of CelebV-HQ

CelebV-HQ consists of 35,666 video clips of 3 to 20 seconds each, involving 15,653 identities, with a total video duration of about 65 hours. For facial attributes, the attribute distribution of CelebV-HQ covers time-invariant (*i.e.*, appearance), time-variant attributes (*i.e.*, action and emotion).

As shown in Table 1, compared to the image datasets that contain facial attribute annotations [45,34], the resolution of CelebV-HQ is more than twice that of CelebA [34], and has a comparable scale to the high-quality dataset, CelebA-HQ [45]. More importantly, CelebV-HQ, as a video dataset, contains not only

appearance attribute annotations, but also action and emotion attribute annotations, which make it contains richer information than image datasets. Other than the diverse annotations, compared to the recent in-the-wild video datasets (CelebV [76], FaceForensics [56], VoxCeleb [51] and VoxCeleb2 [9]), CelebV-HQ has a much higher resolution. Specifically, VoxCeleb2 [9] and MEAD [72], as two representative face video datasets, are the largest audiovisual video face datasets under in-the-wild and lab-controlled environments respectively. Although the data volume of VoxCeleb2 [9] and MEAD [72] is relatively large, the videos on these two datasets are homogeneous and in limited distributions. The videos on VoxCeleb2 are mainly talking face, while MEAD was collected in a constrained laboratory environment. In contrast, CelebV-HQ is collected in real-world scenarios with a diverse corpus, making it more natural and rich in the distribution of attributes.

We start our analysis of CelebV-HQ with the attribute distribution.
1) CelebV-HQ contains a total of 40 appearance attributes, as shown in Fig. 2 (a), of which 10 attributes account for more than 20% each, while there are more than 10 attributes accounting for about 10% each. Meanwhile, the overall attribute distribution has a long tail, with 10 attributes accounting for less than 3% each. We compare the hair colors separately, as they are mutually exclusive. From Fig. 2 (b), the distribution in hair color is even, and there are no significant deviations. 2) There are diverse action attributes in CelebV-HQ as shown in Fig. 2 (c). The common actions, such as "talk", "smile", and "head wagging", account for over 20% each. About 20 uncommon actions, such as "yawn", "cough" and "sneeze", account for less than 1% each. This result is in line with our expectation that these uncommon attributes remain open challenges for the academic community. 3) The proportion of emotion attributes also varies as shown in Fig. 2 (d), with "neutral" accounting for the largest proportion, followed by "happiness" and "sadness" emotions. Unlike the data collected in the laboratory, we do not strictly control the proportion of each attribute, so the overall distribution is more in line with the natural distribution. Overall, the CelebV-HQ is a real-world dataset with diverse facial attributes in a *natural distribution*, bringing new opportunities and challenges to the community.

4.2 Comparison with Image Dataset

Due to CelebV-HQ can be considered as a video version of CelebA-HQ [34] which is a commonly used dataset and its facial attributes annotation is successful in many works [60,66,80,21,11,36,37,8]. We argue that a face video dataset that has similar distribution with CelebA-HQ would also effective.
To show a reasonable distribution of CelebV-HQ, we compare the proposed CelebV-HQ with CelebA-HQ [34] in face attribute aspects such as age, ethnicity, and face shape. Face shape comparisons are provided in the supplementary materials. These factors reflect the basic face information in terms of facial appearance and geometry. Since ethnicity and age attributes are not explicitly labeled, we estimate them for both datasets using an off-the-shelf facial attribute analysis framework [59].

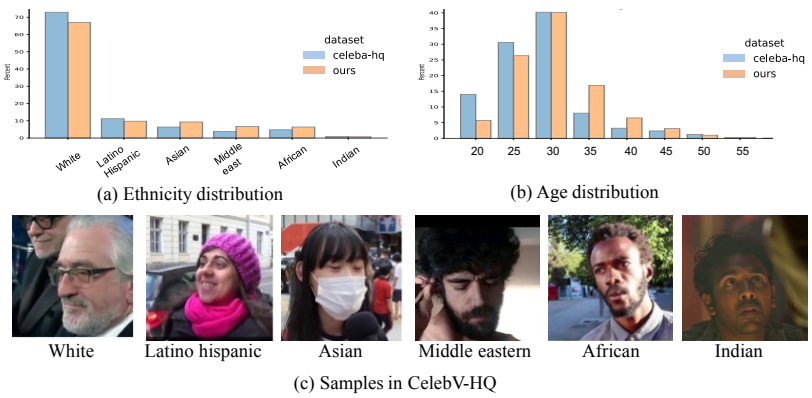


Fig. 3: **Distributions of age and ethnicity compared with CelebA-HQ [34].** (a) and (b) show that CelebV-HQ has a similar distribution compared to CelebA-HQ [34].

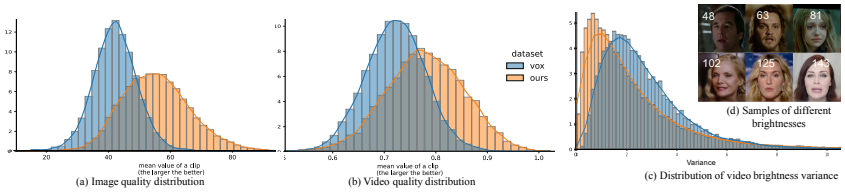
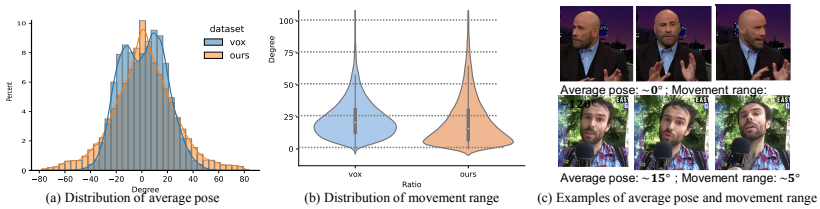


Fig. 4: **Distributions of image and video quality, and brightness variance.** (a) Image quality and (b) video quality are measured by BRISQUE [49] and VSFA [40], the higher score, the better quality. (c) The video brightness is measured by [1], the low variance reflects the more stable in brightness aspect. (d) Samples at different brightness, with brightness values in the upper right corner.

Age distribution. We evaluate whether the dataset is biased towards certain age groups. From Fig. 3 (a), we can see that the age in CelebA-HQ is mainly distributed below 35 years old, while the age distribution of CelebV-HQ is smoother.
Ethnic Distribution. The ethnic distribution roughly reflects the data distribution in terms of geography. As shown in Fig. 3 (b), CelebV-HQ achieves a distribution close to CelebA-HQ [34], and has a more even distribution in Latino Hispanic, Asian, Middle-eastern, and African. As shown in Fig. 3 (c), we show a random sample of each ethnic group in the CelebV-HQ.

4.3 Comparison with Video Dataset

As stated before, VoxCeleb2 [9] is one of the largest in-the-wild face video datasets, and it contains massive face videos, that have contributed to the development of many fields [13,89,16,85,5,88,86,30]. However, CelebV-HQ not only contains speech-based videos, we believe that if it is more diverse and of higher quality than the videos in VoxCeleb2 [9], this can be used to further improve the



(a) Distribution of average pose (b) Distribution of movement range (c) Examples of average pose and movement range

Fig. 5: **Distributions of average head pose and movement range.** There is a wide range of head movement in CelebV-HQ, including both stable videos (less than 20° of movement) and videos with significant movement (from 75° to 100°).

performance of the related models. To demonstrate the superiority of CelebV-HQ and its ability to better support relevant studies, we compare with VoxCeleb2 [9] in terms of the data quality and temporal smoothness. For temporal smoothness, we conduct a comprehensive evaluation in brightness and head pose. We also provide a comparison of the richness and smoothness of the action units in the supplementary material.

Data Quality. We use BRISQUE [49] as a static quality evaluation metric, which is a non-reference evaluation algorithm. For each video clip, we average the BRISQUE [49] value between frames. For the comparison of video quality distributions, we apply the VSFA [40] measurement, a non-reference evaluation method that scores content dependency and temporal memory effects. The distributions are shown in Fig. 4, and CelebV-HQ offers higher quality than VoxCeleb2 [9] at both image and video levels.

Brightness Variation. We also compare video brightness variance distribution with VoxCeleb2 [9]. We first obtain the brightness of each frame and compute the standard deviation in the temporal dimension. The brightness is calculated by averaging the pixels and then converting them to "perceived brightness" [1]. The lower variance of brightness indicates more similar luminance within the video clip, *i.e.*, better brightness uniformity. Fig. 4 (c) shows that CelebV-HQ contains more low variance videos, that demonstrates the brightness change during videos in CelebV-HQ is more stable in the temporal dimension. As shown in Fig. 4 (c), CelebV-HQ contains videos of diverse brightness conditions, we categorized the videos in terms of brightness to further facilitate the usage of CelebV-HQ.

Head Pose Distribution. The head pose distribution is compared in two aspects: the average head pose of a video and the range of head pose movement. These two are used to show the diversity of head poses across the dataset and within the videos, respectively. As stated before, we leverage [75] to detect the head pose in the yaw direction. As shown in Fig. 5 (a), CelebV-HQ is more diverse and smoother than VoxCeleb2 [9] in the overall distribution. From Fig. 5 (b), we see that we have about 75% of the data with movements less than 30°, which means that most of the data are stable, while there are still 25% of movements between 30° and 100°, indicating the overall distribution is diverse. The illustration of average head pose and movement range is shown in Fig. 5 (c).

Fig. 6: **Qualitative results of unconditional video generation.** We present "Full set" and "Subset" settings of MoCoGAN-HD [65] and DIGAN [81]respectively.

5 Evaluation

In this section, we describe our experimental setups and the implementation details of baseline methods. We report the results on state-of-the-art baselines in two typical video generation/editing tasks, *i.e.*, unconditional video generation and video facial attribute editing.

5.1 Unconditional Video Generation

Settings. We employ four unconditional video generation methods, *i.e.*, VideoGPT [78], MoCoGAN-HD [65], DIGAN [81], and StyleGAN-V [63]. We chose these methods based on their performance and code availability. Furthermore, since CelebV-HQ contains action annotations, the models are evaluated under two settings, *i.e.*, the full set of data and the subsets split by different action attributes, *e.g.*, smile. We followed the original authors' setting. To evaluate the model performance, we leverage FVD [69] and FID [22] to access video quality and image quality, respectively.

Results. As shown in Fig. 6, MoCoGAN-HD [65] and DIGAN [81] can generate consistent videos trained on CelebV-HQ. Besides, all methods can successfully produce the desire actions when trained on different subsets of CelebV-HQ with specific attributes. Satisfactory results achieved on the these state-of-the-art methods, demonstrating the effectiveness of CelebV-HQ. More results of different methods are provided in the supplementary materials.

Table 2: **Quantitative results of unconditional video generation.** We evaluate VideoGPT [78], MoCoGAN-HD [65], DIGAN [81], and StyleGAN-V [63] on different datasets and report the FVD and FID scores. "↓" means a lower value is better.

	FaceForensics [56]		Vox [51]		MEAD [72]		CelebV-HQ	
	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)	FVD (↓)	FID (↓)
VideoGPT [78]	185.90	38.19	187.95	65.18	233.12	75.32	177.89	52.95
MoCoGAN-HD [65]	111.80	**7.12**	314.68	**55.98**	245.63	32.54	212.41	21.55
DIGAN [81]	62.50	19.10	201.21	72.21	165.90	43.31	72.98	19.39
StyleGAN-V [63]	**47.41**	9.45	**112.46**	60.44	**93.89**	**31.15**	**69.17**	**17.95**

Table 3: **Quantitative results of video facial attribute editing.** The "Video" version achieves lower FVD scores and comparable FID performance than "Original".

Metrics	**StarGAN-v2 (Gender)**				**MUNIT (Gender)**	
	Original		Video		Original	Video
	Reference	Label	Reference	Label		
FVD (↓)	284.80	258.36	262.01	189.40	219.96	211.45
FID (↓)	**80.61**	65.70	82.99	**55.73**	58.58	**57.01**

Benchmark. We construct a benchmark of unconditional video generation task, for four currently prevalent models (VideoGPT [78], MoCoGAN-HD [65], DIGAN [81], and StyleGAN-V [63]) on 4 face video datasets (FaceForensics [56], Vox [51], MEAD [72] and CelebV-HQ). The benchmark is presented in Table 2. Firstly, it can be observed that the ranking achieved by CelebV-HQ is similar to other prevalent datasets within different methods, which indicates the effectiveness of CelebV-HQ. In addition, the current video generation models [78,65,81,63] obtained good FVD/FID metrics compared to the Vox [51] dataset with similar data size. This illustrates that CelebV-HQ further exploits the potential of the current work, allowing it to generate more diverse and higher quality results. However, CelebV-HQ as a challenging real-world dataset, still has room for community to make improvement.

5.2 Video Facial Attribute Editing

Settings. We employ two representative facial editing baselines, *i.e.*, StarGAN-v2 [8] and MUNIT [26], to explore the potential of CelebV-HQ on video facial attribute editing task. The canonical StarGAN-v2 [8] and MUNIT [26] are designed for image data. We also modify these models by simply adding a vanilla temporal constraint, *i.e.*, estimating the optical flows for adjacent frames in different domains by LiteFlowNet [27] and enforcing L2 Loss between flows. Other losses the original authors proposed remain unchanged. To demonstrate the practical value of our dataset, we select a commonly used appearance attribute, *i.e.*, "Gender", for different baselines.

Results. The baseline methods achieve good results when editing the Gender attribute. The main difference lies in the temporal consistency. In Fig. 7, we observe that the results generated by the original image models are sometimes unstable in the hair area. As reported in Table 3, the "Video" version outperform the

Fig. 7: **Qualitative results of video facial attribute editing.** Results of "Original" tend to have a jittering in the hair area, while results of "Video" are more stable.

"Original" one with respect to the FVD metric in all cases (highlighted in blue), with comparable FID scores. These results indicate that a simple modification using the temporal cues in video dataset can bring performance improvement.

6 Discussion

6.1 Empirical Insights

Some empirical insights are drawn during the construction of CelebV-HQ and the baseline benchmarking. 1) We observe a trend in the growing demand for video facial editing due to the prevalence of short videos [48,28]. However current applications are mainly based on static images [47,28]. Therefore, the research on transforming face editing from images to videos would be an emerging direction. 2) An effective video alignment strategy is important for coherent video generation. In most image generation studies, faces are usually aligned by key points. And the quality might degrade if faces are not aligned. This suggests a new method that can simultaneously retain temporal information and align the face may improve the video face generation.

6.2 Future Work

Finally, we envision the research areas that may benefit from CelebV-HQ.
Video Generation/Editing. CelebV-HQ provides the possibility of improving Video Generation/Editing, such as unconditional face generation [17,55,36,63], text-to-video generation [43,74,23], video facial attributes editing [77], face reenactment [76,62,82], and face swapping [41,87,15]. These tasks rely heavily on the scale and quality of the dataset. CelebV-HQ also contains rich facial annotations, this would allow researcher to go deeper when using this information, *e.g.*, synthesize text description with templates and learning disentanglement of facial attributes.
Neural Rendering. CelebV-HQ has great potential for applications in Neural Rendering. Current tasks, such as novel view synthesis [24,18,52,7,44] and 3d generation [4,52,3,19,14,6], are trained on in-the-wild image datasets [34,36] which lacks facial dynamics to provide natural geometries. CelebV-HQ provides diverse natural facial dynamics and 3D geometries. These features on video modality could not only be further exploited to improve the quality of current models, but also stimulate the emerging of several budding topics, such as Dynamic NeRF [54] and Animatable NeRF [53].
Face Analysis. Face Analysis tasks, such as Attribute Recognition [84,11,33], Action Recognition [71,29], Emotion Recognition [12,39], Forgery Detection [42,20,90], and Multi-modal Recognition [83,50]. These tasks usually require the dataset to have diverse attribute coverage and natural distribution. CelebV-HQ not only meets these requirements, but also could help to transfer previous image tasks to the video version by learning spatio-temporal representations.

7 Conclusion

In this paper, we propose a large-scale, high-quality, and diverse video dataset with rich facial attributes, called CelebV-HQ. CelebV-HQ contains 35,666 video clips involving 15,653 identities, accompanied by 40 appearance attributes, 35 action attributes, and 8 emotion attributes. Through extensive statistical analysis of the dataset, we show the rich diversity of CelebV-HQ in terms of age, ethnicity, brightness, motion smoothness, pose diversity, data quality, *etc.* The effectiveness and future potential of CelebV-HQ are also demonstrated via the unconditional video generation and video facial attribute editing tasks. We finally provide an outlook on the future prospects of CelebV-HQ, which we believe can bring new opportunities and challenges to the academic community. In the future, we are going to maintain a continued evolution of CelebV-HQ, including the scale, quality and annotations.

Acknowledgement. This work is supported by Shanghai AI Laboratory and SenseTime Research. It is also supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Bezryadin, S., Bourov, P., Ilinih, D.: Brightness calculation in digital image processing. In: TDPF (2007)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2018)
3. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: CVPR (2022)
4. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR (2021)
5. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: CVPR (2019)
6. Chen, Y., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Sem2nerf: Converting single-view semantic masks to neural radiance fields. In: ECCV (2022)
7. Cheng, W., Xu, S., Piao, J., Qian, C., Wu, W., Lin, K., Li, H.: Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. arXiv preprint **arxiv:2204.11798** (2022)
8. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR (2020)
9. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
10. Da Xu, L., He, W., Li, S.: Internet of things in industries: A survey. IEEE TII **10**, 2233–2243 (2014)
11. Ding, H., Zhou, H., Zhou, S., Chellappa, R.: A deep cascade network for unaligned face attribute classification. In: AAAI (2018)
12. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors **20**, 592 (2020)
13. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM TOG **37**, 1–11 (2018)
14. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021)
15. Gao, G., Huang, H., Fu, C., Li, Z., He, R.: Information bottleneck disentanglement for identity swapping. In: CVPR (2021)
16. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: CVPR (2021)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
18. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In: ICLR (2021)
19. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: ICCV (2021)
20. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR (2021)
21. Han, H., Jain, A.K., Wang, F., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE TPAMI **40**, 2597–2609 (2017)

22. 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)
23. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint **arXiv:2205.15868** (2022)
24. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: CVPR (2022)
25. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: ECCV Workshop (2008)
26. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
27. Hui, T.W., Loy, C.C.: LiteFlowNet3: Resolving Correspondence Ambiguity for More Accurate Optical Flow Estimation. In: ECCV (2020)
28. Inc., S.: Snapchat. In: https://www.snapchat.com/ (2022)
29. Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A.: Vision-based human action recognition: An overview and real world challenges. Forensic Science International: Digital Investigation **32**, 200901 (2020)
30. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: CVPR (2021)
31. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Deceive D: adaptive pseudo augmentation for GAN training with limited data. In: NeurIPS (2021)
32. Jiang, Y., Huang, Z., Pan, X., Loy, C.C., Liu, Z.: Talk-to-edit: Fine-grained facial editing via dialog. In: ICCV (2021)
33. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: WACV (2021)
34. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
35. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021)
36. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
37. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
38. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
39. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: ICCV (2019)
40. Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: ACM MM (2019)
41. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint **arxiv:1912.13457** (2019)
42. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR (2020)
43. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: AAAI (2018)
44. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: CVPR (2022)

45. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
46. 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)
47. Ltd., F.T.: Faceapp. In: https://www.faceapp.com/ (2022)
48. Ltd., T.P.: Tiktok. In: https://www.tiktok.com (2022)
49. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE TIP **21**, 4695–4708 (2012)
50. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (2020)
51. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
52. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: CVPR (2022)
53. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021)
54. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
55. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint **arxiv:1511.06434** (2015)
56. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint **arxiv:1803.09179** (2018)
57. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: ICCV (2017)
58. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: NeurIPS (2018)
59. Serengil, S.I., Ozpinar, A.: Hyperextended lightface: A facial attribute analysis framework. In: ICEET (2021)
60. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR (2017)
61. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE TPAMI **44**(4), 2004–2018 (2022)
62. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NeurIPS (2019)
63. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: CVPR (2022)
64. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NeurIPS (2014)
65. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. In: ICLR (2020)
66. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM TOG **40**(4), 1–14 (2021)
67. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR (2018)
68. Tzaban, R., Mokady, R., Gal, R., Bermano, A.H., Cohen-Or, D.: Stitch it in time: Gan-based facial editing of real videos. arXiv preprint **arxiv:2201.08361** (2022)

69. 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)
70. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS (2016)
71. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
72. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: ECCV (2020)
73. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: CVPR (2021)
74. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: N\" uwa: Visual synthesis pre-training for neural visual world creation. arXiv preprint **arXiv:2111.12417** (2021)
75. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR (2018)
76. Wu, W., Zhang, Y., Li, C., Qian, C., Loy, C.C.: Reenactgan: Learning to reenact faces via boundary transfer. In: ECCV (2018)
77. Xu, Y., Yin, Y., Jiang, L., Wu, Q., Zheng, C., Loy, C.C., Dai, B., Wu, W.: TransEditor: Transformer-based dual-space GAN for highly controllable facial editing. In: CVPR (2022)
78. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint **arxiv:2104.10157** (2021)
79. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: ICCV (2021)
80. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: ICCV (2021)
81. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. In: ICLR (2021)
82. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: ECCV (2020)
83. Zhang, J., Yin, Z., Chen, P., Nichele, S.: Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion **59**, 103–126 (2020)
84. Zhong, Y., Sullivan, J., Li, H.: Face attribute prediction using off-the-shelf cnn features. In: ICB (2016)
85. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI (2019)
86. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: CVPR (2021)
87. Zhu, H., Fu, C., Wu, Q., Wu, W., Qian, C., He, R.: Aot: Appearance optimal transport based identity swapping for forgery detection. In: NeurIPS (2020)
88. Zhu, H., Huang, H., Li, Y., Zheng, A., He, R.: Arbitrary talking face generation via attentional audio-visual coherence learning. In: IJCAI (2021)
89. Zhu, H., Luo, M.D., Wang, R., Zheng, A.H., He, R.: Deep audio-visual learning: A survey. IJAC **18**, 351–376 (2021)
90. Zhu, X., Wang, H., Fei, H., Lei, Z., Li, S.Z.: Face forgery detection by 3d decomposition. In: CVPR (2021)