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

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) Video Samples (left: general appearance, middle: detailed appearance, right: light conditions) (b) Text Descriptions

Figure 1. **Overview of CelebV-Text.** CelebV-Text contains (a) 70,000 video samples and (b) 1,400,000 text descriptions. Each video sample is annotated with general appearance, detailed appearance, light conditions, action, emotion, and light directions.

Abstract

*Text-driven generation models are flourishing in video generation and editing. However, face-centric text-to-video generation remains a challenge due to the lack of a suitable dataset containing high-quality videos and highly relevant texts. This paper presents **CelebV-Text**, a large-scale, diverse, and high-quality dataset of facial text-video pairs, to facilitate research on facial text-to-video generation tasks. CelebV-Text comprises 70,000 in-the-wild face video clips with diverse visual content, each paired with 20 texts generated using the proposed semi-automatic text generation strategy. The provided texts are of high quality, describing both static and dynamic attributes precisely. The superiority of CelebV-Text over other datasets is demonstrated via comprehensive statistical analysis of the videos, texts, and text-video relevance. The effectiveness and potential of CelebV-Text are further shown through extensive self-evaluation. A benchmark is constructed with representative methods to standardize the evaluation of the facial text-to-video generation task. All data and models are publicly available*[1].

1. Introduction

Text-driven video generation has recently garnered significant attention in the fields of computer vision and computer graphics. By using text as input, video content can be generated and controlled, inspiring numerous applications in both academia and industry [5,34,43,47]. However, text-to-video generation still faces many challenges, particularly in the face-centric scenario where generated video frames often lack quality [18,34,37] or have weak relevance to input texts [2,4,39,67]. We believe that one of the main issues is the absence of a well-suited facial text-video dataset containing high-quality video samples and text descriptions of various attributes highly relevant to videos.

Constructing a high-quality facial text-video dataset poses several challenges, mainly in three aspects. 1) *Data collection.* The quality and quantity of video samples largely determine the quality of generated videos [11, 45, 48, 60]. However, obtaining such a large-scale dataset with high-quality samples while maintaining a natural distribution and smooth video motion is challenging. 2) *Data annotation.* The relevance of text-video pairs needs to be ensured. This requires a comprehensive coverage of text for describing the content and motion appearing in the video, such as light conditions and head movements. 3) *Text gen-*

*Equal contribution.
[1]Project page: https://celebv-text.github.io

eration. Producing diverse and natural texts are non-trivial. Manual text generation is expensive and not scalable. While auto-text generation is easily extensible, it is limited in naturalness.

To overcome the challenges mentioned above, we carefully design a comprehensive data construction pipeline that includes data collection and processing, data annotation, and semi-auto text generation. First, to obtain raw videos, we follow the data collection steps of CelebV-HQ, which has proven to be effective in [66]. We introduce a minor modification to the video processing step to improve the video's smoothness further. Next, to ensure highly relevant text-video pairs, we analyze videos from both temporal dynamics and static content and establish a set of attributes that may or may not change over time. Finally, we propose a semi-auto template-based method to generate texts that are diverse and natural. Our approach leverages the advantages of both auto- and manual-text methods. Specifically, we design a rich variety of grammar templates as [10, 52] to parse annotation and manual texts, which are flexibly combined and modified to achieve high diversity, complexity, and naturalness.

With the proposed pipeline, we create **CelebV-Text**, a Large-Scale Facial Text-Video Dataset, which includes $70,000$ in-the-wild video clips with a resolution of at least 512×512 and $1,400,000$ text descriptions with 20 for each clip. As depicted in Figure 1, CelebV-Text consists of high-quality video samples and text descriptions for realistic face video generation. Each video is annotated with three types of static attributes (40 general appearances, 5 detailed appearances, and 6 light conditions) and three types of dynamic attributes (37 actions, 8 emotions, and 6 light directions). All dynamic attributes are densely annotated with start and end timestamps, while manual-texts are provided for labels that cannot be discretized. Furthermore, we have designed three templates for each attribute type, resulting in a total of 18 templates that can be flexibly combined. All attributes and manual-texts are naturally described in our generated texts.

CelebV-Text surpasses existing face video datasets [11] in terms of resolution (over 2 times higher), number of samples, and more diverse distribution. In addition, the texts in CelebV-Text exhibit higher diversity, complexity, and naturalness than those in text-video datasets [19, 66]. CelebV-Text also shows high relevance of text-video pairs, validated by our text-video retrieval experiments [17]. To further examine the effectiveness and potential of CelebV-Text, we evaluate it on a representative baseline [19] for facial text-to-video generation. Our results show better relevance between generated face videos and texts when compared to a state-of-the-art large-scale pretrained model [26]. Furthermore, we show that a simple modification of [19] with text interpolation can significantly improve temporal coherence. Finally, we present a new benchmark for text-to-video generation to standardize the facial text-to-video generation task, which includes representative models [5, 19] on three text-video datasets.

The main contributions of this work are summarized as follows: 1) We propose CelebV-Text, the first large-scale facial text-video dataset with high-quality videos, as well as rich and highly-relevant texts, to facilitate research in facial text-to-video generation. 2) Comprehensive statistical analyses are conducted to examine video/text quality and diversity, as well as text-video relevance, demonstrating the superiority of CelebV-Text. 3) A series of self-evaluations are performed to demonstrate the effectiveness and potential of CelebV-Text. 4) A new benchmark for text-to-video generation is constructed to promote the standardization of the facial text-to-video generation task.

2. Related Work

Text-to-Video Generation. Text-driven video generation, which involves generating videos from text descriptions, has recently gained significant interest as a challenging task. Mittal *et al*. [43] first introduced this task to generate semantically consistent videos conditioned on encoded captions. Other studies, such as [5, 15, 46], attempt to generate video samples conditioned on encoded text inputs. However, due to the low richness of text descriptions and the small number of data samples, the generated video samples are often at low resolution or lack relevance with the input texts. More recently, several works [19, 26, 27, 55, 57–59] have employed discrete latent codes [16, 54] for more realistic video generation. Some of these works treat videos as a sequence of independent images [19, 26, 58, 59], while Phenaki [55] considers temporal relations between each frame for a more robust video decoding process. Another branch of studies leverage diffusion models for text-to-video generation [20, 24, 25, 51], which require millions or billions of samples to achieve high-quality generation. While text-to-video generation methods are rapidly evolving, they are generally designed for generating general videos. Among these methods, only MMVID [19] has conducted specific experiments with face-centric descriptions. One possible reason for this is that facial text-to-video generation requires more accurate and detailed text descriptions than general tasks. However, there is currently no suitable dataset available that provides such properties for face-centric text-to-video generation.

Multimodal Datasets. Existing multimodal datasets can be categorized into two classes: open-world and closed-world. Open-world datasets [3, 9, 13, 32, 35, 36, 40, 43, 49, 50, 62, 65] are widely used for text-to-image/video generation tasks. Some of them have manual annotations [13, 32, 35, 50, 62] and part of them are directly collected from the Internet, such as subtitles [40, 49]. Closed-world datasets are mostly composed of images or videos collected in constrained environment with corresponding information such as text. CLEVR [28] is a synthetic text-image dataset produced by arranging 3D objects with different shapes under a controlled background. While MUGEN [21] is a video-audio-text dataset that was collected using CoinRun [12] by introducing audio and new interactions. The corresponding text is produced by human annotators and grammar templates.

Multimodal face datasets also exist. Modified MUG [1]

Table 1. **In-the-wild face video dataset comparison.** The symbol "#" indicates the number. The abbreviations "Res.", "Dura.", "App.", "Cond.", "Act.", "Emo.", and "Dir." stand for Resolution, Duration, Appearance, Condition, Action, Emotion, and Direction, respectively. The "half checkmark" denotes that CelebV-HQ consists of action attributes with no timestamp.

Datasets	Meta Information			Attribute Labels						Text	
				Static			Dynamic				
	#Samples	Res.	Dura.	General App.	Detail App.	Light Cond.	Act.	Emo.	Light Dir.	Auto	Manual
CelebV [60]	5	256×256	2hrs								
VoxCeleb2 [11]	150,480	224×224	2442hrs								
CelebV-HQ [66]	35,666	512× 512	68hrs	✓	✗	✗	✓	✓	✗	✗	✗
MM-Vox [19]	19,522	224×224	323hrs	✓	✗	✗	✗	✗	✗	✓	✗
CelebV-Text	70,000	512×512+	279hrs	✓	✓	✓	✓	✓	✓	✓	✓

is a closed-world text-video dataset that contains 1,039 videos with subjects showing different emotions, where the text descriptions are generated from facial emotions using a fixed template [30]. MM-Vox [19] contains 19,522 face videos from VoxCeleb [45], with 36 facial attributes manually labeled following CelebA [38] and text descriptions generated via Probabilistic Context-Free Grammar (PCFG) [61]. However, both datasets only contain language descriptions related to static facial attributes without considering the temporal state change (*i.e.*, emotion or action) presented in the original face videos. Moreover, the limited label annotations restrict the diversity of the text descriptions, making them sub-optimal for studying the text-to-video generation task on the face domain. CelebV-HQ [66] is the latest high-quality face video dataset that covers facial annotations, including appearance, movement, and emotion. However, it only provides discrete labels and timestamps, with no text descriptions.

3. Dataset Construction

In this work, we aim to build a facial text-video dataset, which requires not only large-scale video samples of high quality, but also natural and diverse text descriptions that are highly relevant to videos. To achieve so, we propose an efficient pipeline, as shown in Figure 2, to construct CelebV-Text, including Data Collection & Processing, Data Annotation, and Semi-auto Text Generation.

3.1. Data Collection & Processing

Collection. We follow the same strategy as CelebV-HQ [66] due to its effectiveness in large-scale high-quality data collection. Specifically, we firstly generate a large number of queries, including human names, movie titles, vlogs and so on, to retrieve videos that contain human faces with temporally dynamic state changes and abundant facial attributes. Our data are collected from open world with videos downloaded from online resources. Videos with low resolution ($< 512^2$), low time duration (< 5s), and having appeared in CelebV-HQ are filtered out.

Processing. To sample high-quality and diverse video clips from our raw collections, similar steps are followed as CelebV-HQ [66] with modifications. We first filter out video clips with bounding box regions less than 512^2 rather than resize them to the same resolution. In this way, clips are not upsampled or downsampled hence the video quality would not be affected, which leads to various resolutions of collected videos: 56.4% with $512^2 \sim 1024^2$, and 43.6% for 1024^2+. To reduce the face area noise when the background changes, we further change the video splitting strategy. In addition to our focus on the same human motion [6] and identity [14] present in adjacent frames, we split the video into different clips when the background changes by a toolkit [1].

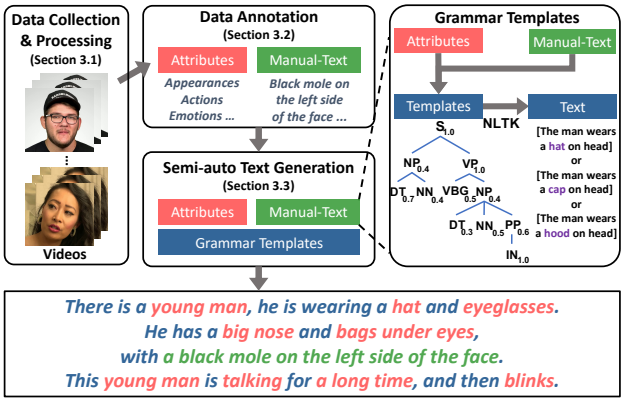


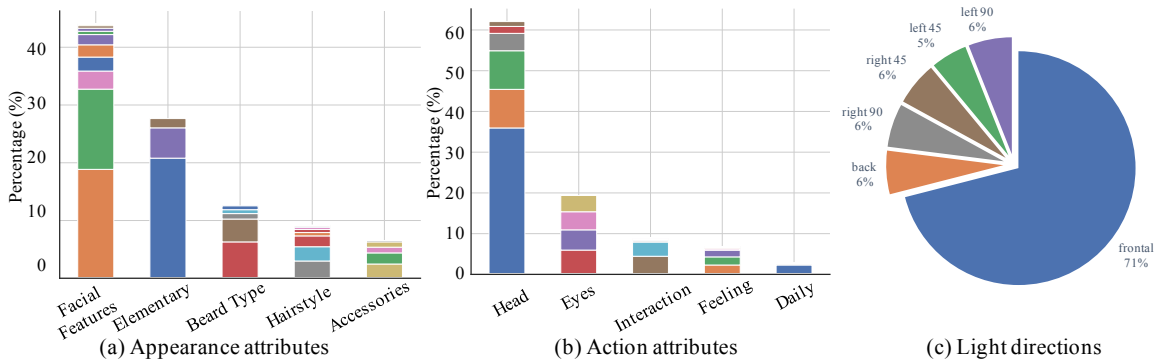
Figure 2. **Pipeline of our dataset construction process.** The pipeline includes data collection & processing, data annotation, and semi-auto text generation.

3.2. Data Annotation

The annotation process is a core part in CelebV-Text construction, which would greatly affect the relevance of text-video pairs, as our designed text templates heavily depend on the annotation results. Here, we first describe how we design attributes, and then give details about the annotation strategy for face videos.

Attributes Design. Temporal dynamic is the key difference between images and videos. However, as shown in Table 1, most face video datasets focus on static attributes where attribute information does not change over time, such as appearance. Dynamic attributes that change over time, such as emotion and face actions, are often neglected. In the fol-

[1] https://github.com/Breakthrough/PySceneDetect



(a) Appearance attributes (b) Action attributes (c) Light directions

Figure 3. **Dataset distribution comparison.** The distributions of appearance attributes, action attributes, and light directions.

lowing, we decouple face videos into static and dynamic categories and details are given as follows.

1) Static. The current dataset [19] only considers static information such as the appearance attribute, which includes 40 classes as CelebA [38]. In contrast, we define static information to include three types of attributes: general appearance, detailed appearance, and light conditions. General appearance attributes follow the same definition as CelebA [38]. Detailed appearance attributes including five classes are proposed for realistic face generation, *i.e.*, scar, mole, freckle, dimple, and one-eyed. We define light conditions in a restricted manner to include light color temperature [22] and brightness [7], with a total of 6 classes.

2) Dynamic. Here, we design three dynamic attributes, *i.e.*, action, emotion, and light directions. For action attributes, we follow CelebV-HQ [66] and expand their action list by two classes, *i.e.*, squint and blink. For emotion attributes, we select the 8 emotion setting in Affectnet [44], including neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise. For light direction attributes, we derive and modify classes from [29] and give 6 light direction classes. Complete lists are given in the Appendix. Moreover, as shown in Table 1, CelebV-HQ [66] is the only dataset giving timestamps of dynamic attributes. Following their idea, we densely annotate all dynamic attributes of CelebV-Text with the start and end time.

Automatic and Manual Annotation. Based on our attributes design, we find that some attributes can be annotated automatically (*e.g.*, appearance) while some need manual annotations (*e.g.*, timestamps of dynamic attributes). Considering the dataset quality and cost of expense, our annotation strategy includes both automatic and manual annotations.

For automatic annotation, we first investigate algorithms and select designed attributes that can be automatically annotated. We then test different algorithms on our dataset and keep those giving annotation accuracy of 85% or higher. This process yields all light condition labels, all appearance labels, and all emotion labels suitable for automatic annotation. Algorithms for different labels we finally chose are reported in the Appendix. Automatic annotation results can be further revised by human workers to improve accuracy in a less costly way.

For manual annotation, we hire and train human workers following [66] to annotate attributes that are filtered out by an automatic annotation process. In this case, we manually annotate dynamic attributes, *i.e.*, action and light directions, to give both class labels and exact timestamps. In addition, it is hard to represent detailed appearance attributes by the discrete label, *e.g.*, the characteristics of scars or moles. We therefore ask annotators to give a natural description for each attribute, describing exact positions relative to face parts. These designs greatly enhance the relevance between the final text and the video.

3.3. Semi-auto Text Generation

Multimodal text-video datasets collect texts via three common methods: subtitles [4, 40, 67], manual-text generation [3, 9, 32, 56, 62], and auto-text generation [5, 21, 27]. However, it is difficult for the individual method to generate texts with high relevance to videos, natural expression, and high diversity. Specifically, although subtitles are easy to obtain, they can pose weakly relevant text-video pairs and introduce noise, making the dataset quality hard to control. Moreover, manual-text generation method is time and cost consuming, as natural language descriptions are required for each video. In this case, increasing the data scale is quite hard as more workers are needed to describe new videos, which does not meet the efficiency and scalability of annotation. Finally, auto-text generation is flexible and scalable, as abundant texts can be simultaneously generated given annotation results of collected videos. However, the diversity, complexity, and naturalness of generated texts can be impacted by the designed grammar templates.

To this end, we propose a semi-auto template-based text generation strategy that combines both manual-text and auto-text generation methods. Specifically, as mentioned in Section 3.2, manual-texts are required to describe detailed appearance attributes. Annotated attribute information is fed into our designed template for auto-text generation.

To make our template as natural as possible, we first ask each annotator to describe 10 different face videos for each attribute. We then analyze the grammar structure (*i.e.*, parse tree banks) along with online corpora following [10, 31], and find the most three common grammar structures for each attribute. Finally, we utilize probabilistic context-free grammar [52, 61] and modify the grammar structures to design our own templates. Texts are generated based on templates with synonym replacement using NLTK [8] to increase our generation diversity. Details of our template de-

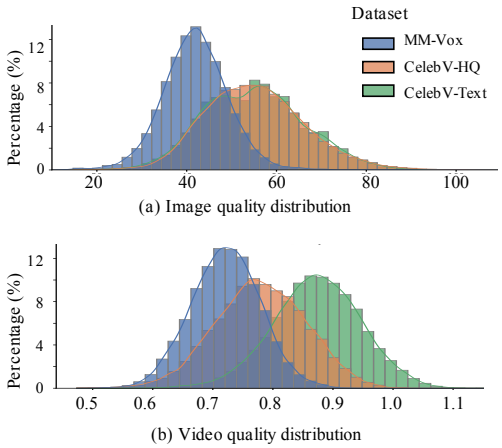


Figure 4. **Dataset quality distribution.** The metrics used are BRISQUE [42] and VSFA [33] respectively.

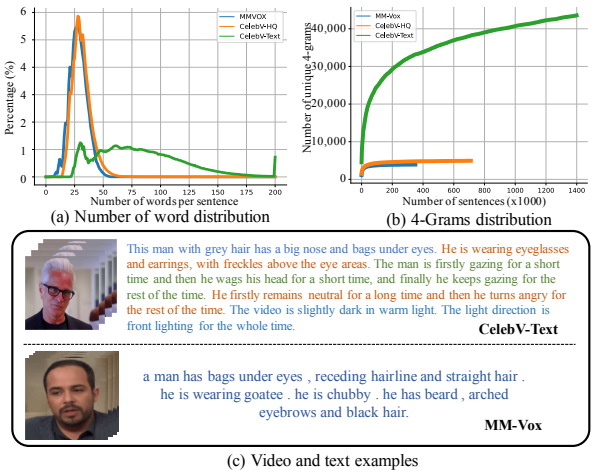


Figure 5. **Text distribution.** CelebV-Text achieves better performance in both 4-gram and number words distribution.

signs are in the Appendix.

4. Statistical Analysis of CelebV-Text

In this section, we compare CelebV-Text with the two most relevant and representative face video datasets [19,66]. We perform a comprehensive analysis of CelebV-Text in terms of video, text, and text-video relevance. To verify the effectiveness of our designed grammar templates, we generate text descriptions for CelebV-HQ based on its attributes for comparison. For simplicity, we use "CelebV-HQ" to denote this variant in the following.

4.1. Video Comparisons

We briefly compare the overall statistics of existing face video datasets [11, 19, 60, 66] in Table 1. As reported, CelebV-Text contains 70,000 video clips with a total duration of around 279 hours. Each video is accompanied by 20 sentences describing all 6 designed attributes. Compared to CelebV [60], CelebV-Text has a larger scale and higher resolution. Although VoxCeleb2 [11] has more samples than CelebV-Text, its video distribution is limited as most videos are mainly talking faces. Moreover, video samples of both CelebV-HQ [66] and CelebV-Text are collected in open-world with diverse queries so that they are rich in distribution, while CelebV-Text has about 2 times video data, more video attributes, and highly relevant text descriptions. Finally, compared to the only existing facial text-video dataset MM-Vox [19], CelebV-Text overpasses MM-Vox in terms of scale and quality.

Attributes Distribution. In order to better present the distribution of different attributes in CelebV-Text, we pick and divide general appearance, action, and light direction attributes into groups. More distributions and division designs are provided in the Appendix. Specifically, all 40 general appearance classes are divided into 5 groups shown in Figure 3 (a). Facial features (*e.g.*, double chin, big nose, and oval face) account for the most portion around 45%. The elementary group is twice large than the beard type, accounting for around 25% and 12%, respectively. Fewer samples are located to the hairstyle and accessories groups, taking around 10% and 8%, respectively. Besides, action attributes are divided into 5 groups in Figure 3 (b), where it is clear that head-related actions account for the largest portion of around 60%, followed by eyes-related actions of around 20%. The interaction group (*e.g.*, eat), feeling group (*e.g.*, smile), and daily group (*e.g.*, sleep) account for around 9%, 7%, and 4%, respectively. Finally, for light directions (Figure 3 (c)), most samples contain the front lighting and the remaining ones are evenly distributed.

Video Quality Distribution. We follow [66] to analyze the quality of our collected videos. To demonstrate the superiority of CelebV-Text, we compare with MM-Vox [19] and CelebV-HQ [66], where mean BRISQUE [42] and VSFA [33] are used to evaluate the image and video quality, respectively. Image quality of all datasets is shown in Figure 4 (a), where CelebV-Text and CelebV-HQ achieve comparable quality, higher than MM-Vox by a large margin. Video quality of all datasets is shown in Figure 4 (b), where CelebV-Text has the best quality, which is due to the effect of the video split method mentioned in Section 3.1, alleviating the discontinuity during background transitions.

4.2. Text Comparisons

In addition to a large number of video samples, text descriptions of CelebV-Text are longer and more detailed than those in MM-Vox [19] and CelebV-HQ [66] (see Figure 5 (a)), where the average text length of MM-Vox, CelebV-HQ, and CelebV-Text are 28.39, 31.06, and 67.15. Distributions of Celeb-HQ and MM-Vox are close, but there are more words in CelebV-Text to describe a video due to the comprehensive annotation.

To validate the linguistic diversity of the generated texts, comparisons are conducted among the three datasets following [56]. Specifically, we report the unique part-of-speech (POS) tags (*i.e.*, verb, noun, adjective, and adverb) of the three datasets in Table 3. Obviously, due to our comprehensively designed attribute list and the number of templates, CelebV-Text presents a wider variety of text styles, covering a broader range of face attributes that are static or

Table 2. **Multimodal retrieval results.** Clip2Video [17] is leveraged to measure the text-video relevance via retrieval experiments. Bold values indicate the best results, underlined ones indicate the second best.

		Text ⇒ Video					Video ⇒ Text				
Description	Dataset	R@1(↑)	R@5(↑)	R@10(↑)	MdR(↓)	MnR(↓)	R@1(↑)	R@5(↑)	R@10(↑)	MdR(↓)	MnR(↓)
(a) App.	MM-Vox [19]	1.5	9.0	15.7	52.0	68.8	2.0	9.2	14.6	43.0	57.8
	CelebV-HQ [66]	5.9	19.2	29.7	27.0	52.2	7.2	20.7	32.4	27.0	46.9
	CelebV-Text	6.1	21.3	35.5	26.3	49.1	7.4	20.7	29.9	26.6	48.3
(b) App.+Emo.	CelebV-HQ [66]	6.5	20.1	30.8	**25.0**	48.0	7.9	25.5	**38.8**	17.0	37.0
	CelebV-Text	6.6	23.4	37.1	26.0	47.6	**8.1**	27.2	34.7	18.2	38.3
(c) App.+Emo.+Act.	CelebV-Text	**6.9**	**24.1**	**39.2**	25.8	**46.7**	8.0	**27.6**	37.1	**16.7**	**36.1**

Table 3. **Number of unique POS tags.** The numbers of unique POS tags for MM-Vox, CelebV-HQ, and CelebV-Text.

Dataset	#Verb	#Adj.	#Noun	#Adv.
MM-Vox [19]	5	20	38	0
CelebV-HQ [66]	10	24	50	6
CelebV-Text	**96**	**78**	**174**	**24**

dynamic in the temporal domain.

In addition, we further examine the naturalness and complexity of our texts compared to MM-Vox, where we modify [63] to calculate the type-token vocabulary curve for all captions. As shown in Figure 5 (b) where unique 4-grams are selected as the types [56], it is evident that due to our grammar structures and synonym replacement, the linguistic naturalness (vocabulary use) and complexity (vocabulary size) of our CelebV-Text are much better. Please refer to Appendix for more n-grams results.

4.3. Text-Video Relevance

To quantitatively validate our text-video relevance, we conduct text-video retrieval tasks on three datasets: MM-Vox [19], CelebV-HQ [66], and CelebV-Text. Rather than use conventional frame-wise clip score as most works [24, 51, 55], we follow [17] to compute feature similarities between texts and videos with the consideration of temporal dynamics, which reflects accurate multimodal interactions across the two modalities. Recall at rank K (R@K), median rank (MdR), and mean rank (MnR) [17, 41, 64] are used as evaluation metrics, where the higher R@K, the lower median rank and mean rank indicate better performance.

We first examine the performance given texts with descriptions of general appearance in Table 2 (a). Results of CelebV-HQ and CelebV-Text are both better than MM-Vox for two retrieval tasks, which indicates our designed templates can produce texts more relevant to videos than MM-Vox. We further add descriptions about dynamic emotion changes to CelebV-HQ and CelebV-Text in Table 2 (b). Similar results are achieved in both datasets, which reflects that our annotation accuracy on static appearance attributes is as good as CelebV-HQ. Finally, we append action descriptions to CelebV-Text in Table 2 (c), which achieves the best performance on most metrics, verifying the relevance between our generated texts and video samples.

5. Experiment

In this section, we first conduct facial text-to-video generation to validate the effectiveness of our CelebV-Text dataset. We then benchmark representative approaches on facial text-to-video generation task.

5.1. High-relevance Text-to-Video Generation

To show the benefits brought by our text descriptions which depict both static and dynamic attributes, we conduct experiments to show the effectiveness of CelebV-Text. Experiments are mainly based on a recent open-sourced state-of-the-art method, MMVID [19], and compared with CogVideo[2] [26], which is a large-scale pretrained text-to-video model, trained on millions of text-image/video pairs.

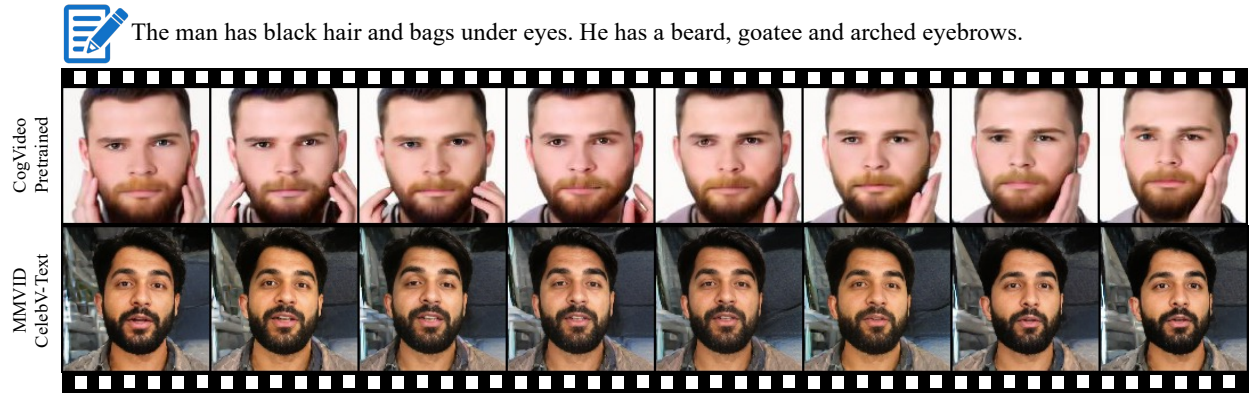
Static Face Video Generation. To validate the effectiveness of our facial text-video dataset in static attributes, we use the models stated above to generate videos conditioned on general appearance, face details, and light conditions descriptions, respectively. Specifically, we first train MMVID [19] from scratch solely on CelebV-Text. We then generate 3 input texts including individual descriptions of each of the static attributes. Generated texts are fed into both MMVID [19] and CogVideo [26] and corresponding video outputs are examined.

Visualization results of general appearance are shown in Figure 6 (a), which prove the effectiveness of our dataset. We observe that although CogVideo can output the face video given a text description, the text-video pair is not quite relevant, such as "bags under eyes" and "wavy hair". However, MMVID [19] produces videos with high relevance to input texts, containing all attributes described in the text. More results are shown in the Appendix.

Dynamic Face Video Generation. We follow the above experimental setting to validate the effectiveness of our dataset with dynamic attribute changes (*i.e.*, emotion, action and light direction). Due to the difficulty in modelling state change [27, 51], we follow [5] to apply test-time interpolation to MMVID [19], named MMVID-interp, to improve the text encoding and better understand the dynamics. Details of our modification are shown in the Appendix.

In Figure 6 (b), we observe that CogVideo fails to reflect the temporal change described in the input text, *i.e.*, smile

[2]We choose CogVideo [26] as the representative large-scale model for comparison, since the inference code and pretrained models of other large-scale methods (*e.g.*, CogVideo [26], Phenaki [55], Imagen Video [24], and Make-A-Video [51]) are not public.



(a) Static - General Appearance

(b) Dynamic - Action

Figure 6. **Qualitative results of facial text-to-video generation.** The generated samples are given texts describing (a) the static attribute and (b) dynamic attribute.

Table 4. **Benchmark of text-to-video generation on different datasets.** ↓ means a lower value is better and ↑ means the opposite.

(a) Quantitative results on general appearance descriptions.

Dataset	Method	FVD(↓)	FID(↓)	CLIPSIM(↑)
MM-Vox [19]	TFGAN [5]	502.28 ± 1.66	760.24 ± 16.01	0.165 ± 0.022
	MMVID [19]	**65.79 ± 1.81**	**38.81 ± 3.66**	**0.170 ± 0.020**
CelebV-HQ [66]	TFGAN [5]	428.04 ± 1.76	616.24 ± 17.45	0.168 ± 0.021
	MMVID [19]	**73.65 ± 1.43**	**63.86 ± 3.66**	**0.172 ± 0.019**
CelebV-Text	TFGAN [5]	403.04 ± 1.34	589.24 ± 16.46	0.177 ± 0.012
	MMVID [19]	**66.69 ± 1.35**	**58.70 ± 4.67**	**0.198 ± 0.014**

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

Dataset	Method	FVD(↓)	FID(↓)	CLIPSIM(↑)
CelebV-Text **App.+Emo.**	TFGAN [5]	442.30 ± 2.56	623.17 ± 18.88	0.158 ± 0.024
	MMVID [19]	82.78 ± 1.47	61.58 ± 3.99	0.176 ± 0.008
	MMVID-interp	**72.87 ± 1.23**	**41.57 ± 3.56**	**0.182 ± 0.010**
CelebV-Text **App.+Act.**	TFGAN [5]	571.34 ± 4.54	784.93 ± 20.13	0.154 ± 0.028
	MMVID [19]	109.25 ± 2.11	82.55 ± 4.37	0.174 ± 0.019
	MMVID-interp	**80.81 ± 2.55**	**70.88 ± 4.77**	**0.176 ± 0.020**

→ turn. However, both MMVID [19] and MMVID-interp trained on CelebV-Text can successfully model the dynamic attribute changes, which demonstrates the effectiveness of our dataset. In addition, we find that MMVID [19] cannot preserve some attributes well (*e.g.*, earrings), while MMVID-interp can stabilize the sampling process, validating the effectiveness of our modification. More results are shown in the Appendix.

Note that CogVideo [26] has a much larger model size (∼ 100 times larger than MMVID [19]) and is trained on much large text-video data (∼ 75 times larger than CelebV-Text). However, video samples produced by CogVideo [26] shown in Figure 6 are of a lower quality than the ones by MMVID [19] trained solely on CelebV-Text, where generated faces are not in a high relevance to input texts, demonstrating the effectiveness of our facial text-video dataset.

5.2. Benchmark on Facial Text-to-Video Generation

As the domain of text-to-video generation is currently thriving, there exists only one benchmark in the face domain, MM-Vox [19]. We expand [19] and construct a benchmark of facial text-to-video generation tasks on three datasets: MM-Vox [19], CelebV-HQ [66] with texts generated by our templates, and CelebV-Text. We choose two representative methods[3], TFGAN [5] and MMVID [19], to evaluate their performances on all datasets.

[3]Other methods, *e.g.*, CogVideo [26], Phenaki [55], Imagen Video [24], and Make-A-Video [51] are not included since their training codes are not

Figure 7. **Qualitative results on three facial text-video datasets.** Red and yellow regions indicate the missing of "bags under eyes" and the existence of "wavy hair" and "bags under eyes".

Quantitative Results. For thorough benchmark construction, we evaluate baseline methods given variant texts including static and dynamic attributes. We use FVD [53] (temporal consistency), FID [23] (individual frame quality), and CLIPSIM [57] (text-video relevance) as evaluation metrics following [19] and report detailed results for appearance, action, and emotion in Table 4. Evaluation steps are repeated over ten runs with mean values and standard errors reported as well. All other values are shown in the Appendix. It can be seen from Table 4 that MMVID [19] obtains good FVD/FID/CLIPSIM metrics over TFGAN [5] which fails to generate reasonable video outputs. In addition, when input texts contain descriptions about a dynamic state change in the temporal domain, the generated video quality by MMVID [19] decreases, which encourages future methods to focus more on cross-modal understanding and consistent video generation. Moreover, the performance of MMVID-interp is better than MMVID [19] on all metrics, validating the effectiveness of our modification mentioned in Section 5.1. Due to challenges posed by our dataset and text-to-video generation task, there is still considerable room to improve.

Qualitative Results. Video samples generated from MMVID [19] trained on different datasets are shown in Figure 7, where all video frames are of 128^2. We can see that video samples generated by MMVID [19] trained on different datasets are of high quality with temporal consistency. However, MMVID [19] trained on MM-Vox [19] can sometimes fail to generate attributes mentioned in the input texts. More generated video samples with dynamic attribute changes are shown in the Appendix.

6. Discussion

We have proposed CelebV-Text, a large-scale, high-quality, and diverse facial text-video dataset with static and dynamic attributes. CelebV-Text contains $70,000$ video clips, each of which is accompanied by 20 individual sentences describing both static and dynamic factors. Through extensive statistical analysis and experiments, we have demonstrated the superiority and effectiveness of CelebV-Text. In the future, we plan to further enlarge CelebV-Text in both scale and diversity. We may further explore several new tasks based on CelebV-Text, such as fine-grained control of video face, adaptation of general pretrained models to the face domain, and text-driven 3D-aware facial video generation.

Ethical Consideration. CelebV-Text is intended for research purposes only. While the raw videos will not be released, the data annotations, links to raw videos, and data processing tools will be made available after undergoing a rigorous legality check procedure at our institution. It is worth noting that our data annotation does not include any personal biometric information such as identity. Only generic attribute information such as gender, hair color, and motion is annotated. Additionally, synthetic videos generated in this work do not exhibit bias or certain biometric information (*e.g.*, big lips or big nose), alleviating ethical concerns. CelebV-Text may be used for deepfakes, but it can also be used for forgery detection tasks to prevent such issues. We will try our best to control the application and acquisition procedure of CelebV-Text to avoid potential misuse and abuse. In the future, we plan to use synthetic face generation frameworks to generate synthetic face videos to address the ethical shortcomings of existing real-world face video datasets.

Acknowledgement. CelebV-Text is developed under OpenXDLab – an open platform for X-Dimension high-quality data. This study is supported by the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, in-kind contribution from the industry partner(s), and the MOE AcRF Tier 1 (RG16/21).

public so far.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *WIAMIS*, 2010. 2

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 1

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2, 4

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 4

[5] 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. 1, 2, 4, 6, 7, 8

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*. IEEE, 2016. 3

[7] Sergey Bezryadin, Pavel Bourov, and Dmitry Ilinih. Brightness calculation in digital image processing. In *TDPF*, 2007. 4

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 4

[9] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 2, 4

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

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1, 2, 3, 5

[12] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *ICML*, 2019. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. In *IJCV*, 2022. 2

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 3

[15] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 2

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[17] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 6

[18] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, 2018. 1

[19] 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, 3, 4, 5, 6, 7, 8

[20] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 2

[21] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Isabelle Hu, and Devi Parikh. Mugen: A playground for video-audio-text multimodal understanding and generation. *arXiv preprint arXiv:2204.08058*, 2022. 2, 4

[22] 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. 4

[23] 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. 8

[24] 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. 2, 6, 7

[25] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[26] 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. 2, 6, 7

[27] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022. 2, 4, 6

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2

[29] Peter Kán and Hannes Kafumann. Deeplight: light source estimation for augmented reality using deep learning. *The Visual Computer*, 2019. 4

[30] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. In *IEEE Access*, 2020. 3

[31] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *ACL*, 2003. 4

[32] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2, 4

[33] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM MM*, pages 2351–2359, 2019. 5

[34] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[36] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework

for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 2

[37] Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. Cross-modal dual learning for sentence-to-video generation. In *ACM MM*, 2019. 1

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3, 4

[39] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *ACL*, 2015. 1

[40] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *CVPR*, 2019. 2, 4

[41] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, 2018. 6

[42] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. In *TIP*, 2012. 5

[43] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *ACM MM*, 2017. 1, 2

[44] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. In *TAC*, 2017. 4

[45] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 1, 3

[46] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *ACM MM*, 2017. 2

[47] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 1

[48] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 1

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[51] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 6, 7

[52] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020. 2, 4

[53] 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. 8

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 2

[55] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2, 6, 7

[56] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 4, 5, 6

[57] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2, 8

[58] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. In *NeurIPS*, 2022. 2

[59] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 2

[60] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 1, 3, 5

[61] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021. 3, 4

[62] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 4

[63] Gilbert Youmans. Measuring lexical style and competence: The type-token vocabulary curve. *Style*, 1990. 6

[64] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018. 6

[65] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2

[66] 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. 2, 3, 4, 5, 6, 7

[67] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 1, 4