Talk2Face: A Unified Sequence-based Framework for Diverse Face Generation and Analysis Tasks

Yudong Li
liyudong2021@email.szu.edu.cn
School of Computer Science and Software Engineering, Shenzhen University
Shenzhen Institute of Artificial Intelligence of Robotics of Society

Xianxu Hou
hxianxu@gmail.com
National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
Shenzhen Institute of Artificial Intelligence of Robotics of Society

Zhe Zhao
nlpzhezhao@tencent.com
Tencent AI Lab

Linlin Shen*
llshen@szu.edu.cn
School of Computer Science and Software Engineering, Shenzhen University
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University

Xuefeng Yang
ryanxfyang@tencent.com
Tencent AI Lab

Kimmo Yan
kimmoyan@tencent.com
Tencent AI Lab

ABSTRACT

Facial analysis is an important domain in computer vision and has received extensive research attention. For numerous downstream tasks with different input/output formats and modalities, existing methods usually design task-specific architectures and train them using face datasets collected in the particular task domain. In this work, we proposed a single model, Talk2Face, to simultaneously tackle a large number of face generation and analysis tasks, e.g. text guided face synthesis, face captioning and age estimation. Specifically, we cast different tasks into a sequence-to-sequence format with the same architecture, parameters and objectives. While text and facial images are tokenized to sequences, the annotation labels of faces for different tasks are also converted to natural languages for unified representation. We collect a set of 2.3M face-text pairs from available datasets across different tasks, to train the proposed model. Uniform templates are then designed to enable the model to perform different downstream tasks, according to the task context and target. Experiments on different tasks show that our model achieves better face generation and caption performances than SOTA approaches. On age estimation and multi-attribute classification, our model reaches competitive performance with those models specially designed and trained for these particular tasks. In practice, our model is much easier to be deployed to different facial analysis related tasks. Code and dataset will be available at https://github.com/ydli-ai/Talk2Face.

CCS CONCEPTS

• **Computing methodologies** → *Natural language generation*; *Computer vision tasks.*

KEYWORDS

cross-modal generation, text-to-face synthesis, face captioning

ACM Reference Format:
Yudong Li, Xianxu Hou, Zhe Zhao, Linlin Shen, Xuefeng Yang, and Kimmo Yan. 2022. Talk2Face: A Unified Sequence-based Framework for Diverse Face Generation and Analysis Tasks. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3548205

1 INTRODUCTION

In the past few years, there are many research and development of automated facial analysis systems such as expression recognition [31, 43, 49], age estimation [4, 5] and text-guided generation [52, 62]. Existing works mainly use supervised learning with large-scale annotated data to achieve state-of-the-art results. In practice, each task requires independent steps of data collection, data annotation, model design, and training. Due to the labor-intensive process, it's very difficult to transfer knowledge across tasks. Previously, multi-task learning (MTL) on several interconnected downstream tasks has been a common scheme for face domain knowledge transfer. Most works [23, 39, 55] use a shared backbone network with task-specific output layers for similar or partially overlapping tasks. However, these MTL frameworks can only handle limited classification tasks. As additional layers are required to learn new tasks, the cross-task generalization capability is limited. In addition, these methods cannot simultaneously deal with the face discriminative and generative tasks.

Recently, inspired by the unified transformer-based models in natural language processing (NLP) [3, 22, 25, 37], multimodal variants [9, 35, 54] have achieved promising performance on many

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548205

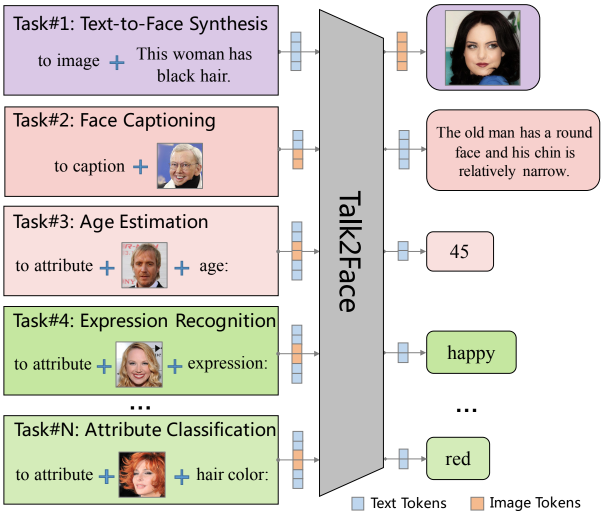

Task#1: Text-to-Face Synthesis
to image
This woman has black hair.
Task#2: Face Captioning
to caption
Task#3: Age Estimation
to attribute
age:
Task#4: Expression Recognition
to attribute
expression:
...
Task#N: Attribute Classification
to attribute
hair color:
Talk2Face
The old man has a round face and his chin is relatively narrow.
45
happy
...
red
Text Tokens
Image Tokens

Figure 1: Our approach represents different face-related tasks with a unified sequence-to-sequence model and uses it for text/face generation.

computer vision (CV) tasks. In addition to the advantage of transformer in architecture, these unified approaches leverage general knowledge to alleviate the reliance on downstream task data. However, in the face domain, unified method is still unexplored. Part of the reason is that face datasets are typically annotated for specific tasks, and there is no method to learn generic face knowledge from datasets of various formats and modalities.

To alleviate above limitations, in this work, we present *Talk2Face*, a unified text-and-face generative model that that explicitly shares knowledge among different tasks. We propose a natural language supervised framework that unifies heterogeneous face related vision tasks into a single sequence-to-sequence format and enables full parameters sharing across different tasks. In essence, we cast different tasks as the same sequence modeling task conditioned on prompts, with the shared model architecture and loss function.

To provide a databank for learning face domain generic knowledge, we create a large-scale dataset of 2.3 million face-text pairs. We collect face datasets from different supervision (e.g., age, race, expression, etc.) and setup protocols to convert their labels into text descriptions. Based on this unified dataset, we train *Talk2Face* using the sequence modeling objective, i.e., feed text to produce face or feed face to produce text.

In the inference phase, downstream tasks are represented with the same format as the training to bridge the gap between generic training and task-specific inference. Specifically, we use templates to transform downstream tasks as sequence-to-sequence format. As shown in Figure 1, text-guided face synthesis can be achieved by feeding a text followed by a prompt "to image", and then the face is generated via autoregressive decoding. Similarly, for the face captioning task, an image and a prompt "to caption" are fed to the model, and a corresponding text is generated. For facial analysis tasks, the model directly generates answers in text format. In contrast to existing methods that separately model different tasks, our model uses the same sequence modeling head to generatively perform all tasks without fine-tuning. Additionally, for learning a new task, generic face knowledge can be readily leveraged by reformulating its input and output.

To assess the performance of this unified and cross-modal training, we evaluate Talk2Face on 6 downstream tasks, including text-guided face synthesis on Multi-Modal CelebA-HQ [51], face image caption on CelebA-text [44], age estimation on CACD [7], race classification on FairFace [20], expression classification on AffectNet [32] and multi-attribute classification on CelebA [29]. For text-guided face synthesis and face image captioning tasks, our unified approach outperforms recent state-of-the-art methods. While the performance of our model on age estimation, expression recognition and race classification is currently not as good as those models specially designed and trained on these particular tasks, our model is general, trained on a single objective and does not require any fine-tuning. We also find that downstream tasks can benefit from generic face knowledge. For example, our model learns fine-grained age information from general training and therefore can control the age of faces generated in text-guided face synthesis, which is difficult to achieve for the previous literature works. Our contributions are summarized as follows:

- We propose the first face knowledge learning framework to allow explicit knowledge transfer among different face-ralated vision tasks.
- We create a large-scale face dataset containing 2.3 million paired text and face images, for general face domain knowledge learning.
- Based on the large scale face and text pairs, a transformer based generative model, Talk2Face, is trained and directly applied to different downstream tasks like text guided face synthesis, face captioning and facial analysis such as age estimation, expression recognition and race classification.

2 RELATED WORK

Text-to-Image Generation. In the past few years, there have been a variety of GAN-based text-to-image synthesis approaches. Reed et al. [40] takes the noise and sentence embedding as the input for a one-stage GAN framework to synthesize 64×64 images. StackGAN [56] and StackGAN++ [57] further design a multi-stage framework to generate 256×256 images. AttnGAN [53] implements a multi-stage module StackGAN++ by using attention mechanism. ControlGAN [26] uses a word-level spatial and channel-wise attention-driven generator to correlate words with image regions.

In face domain, Xia et al. [51] adopt StyleGAN [21] as backbone to synthesize high-resolution of 1024×1024 face images. Sun et al. [44] use semantic embedding and attention network to generate face images based on multiple captions input. Zhou and Shimada [62] use pre-trained BERT model to obtain text embedding and StyleGAN encoder for learning latent representations to perform text-to-face generation on a small dataset. TextFace [17] encode text description into the latent space of a pre-trained StyleGAN to guide face images generation.

Recently, transformer-based autoregressive language models have greatly improved the performance of various text generation

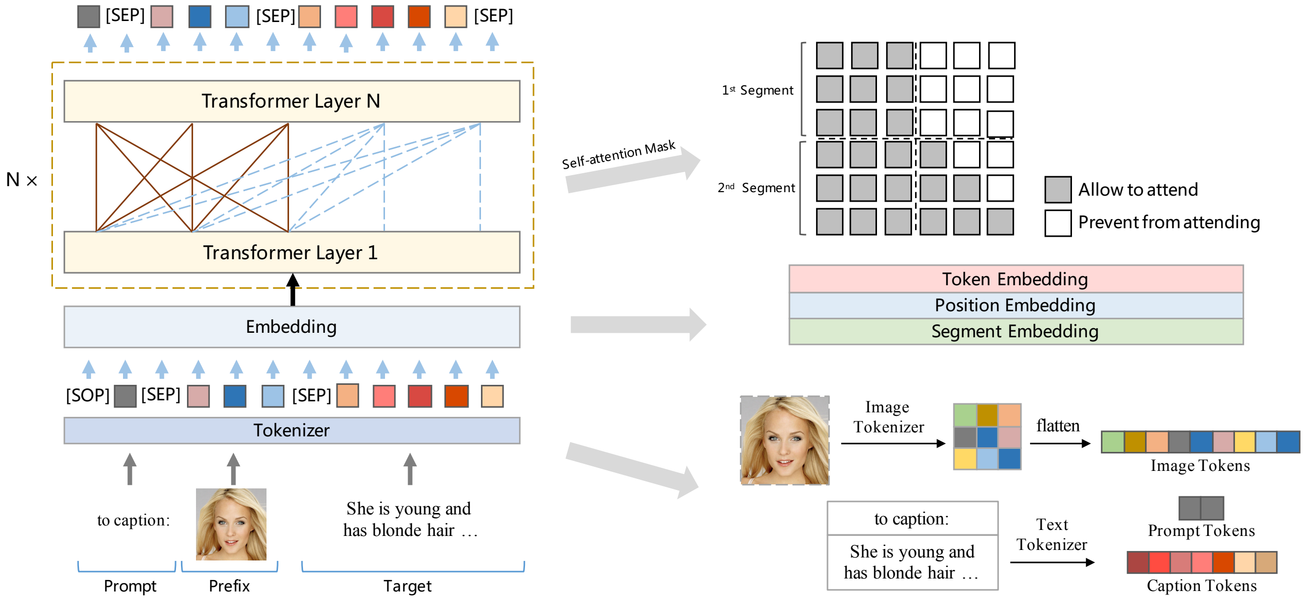


Figure 2: Overview of our approach. Text and faces are represented as discrete sequences conditioned on prompts. Input sequences are left-shifted for language modeling objective. We use self-attention mask to control the access to context for each token.

tasks, such as machine translation and text summarization. Utilizing image quantization to represent images as discrete sequences, transformer-based methods have achieved promising results on high-resolution image synthesis [13] and text-to-image synthesis [10, 38].

Image-to-Text Generation. A typical task of image-to-text generation is image captioning, which converts an image into text that describes the image's content. Existing methods mainly adopt encoder-decoder architectures, where an image is input to the encoder and text is generated in the decoder. Vinyals et al. [47] use CNN to produce the image representation and use the last hidden layer as an input to the LSTM decoder to generate sentences. Zhou et al. [61] use a shared multi-layer transformer network for both encoding and decoding, which fuses the information from both modalities.

Image-to-text generation is further used to unify the labels of different downstream tasks with natural language. Compared to the output bound with the linear layer, natural language supervision provides wider supervision for visual concepts. VL-T5 [9] handles visual question answering, visual grounding, and image-text matching tasks by generating labels in text format, for which a special text token is used to assign a specific region in the image. PIX2SEQ[8] tackles object detection task by converting bounding boxes and class labels into a sequence of discrete tokens. The model receives pixel inputs and generates the target sequence.

Compared to the prosperity of image-captioning works for natural images, the face-to-text related works is very limited. As face attributes are fine-grained and closely related, it might be more difficult to describe them than different objects, which requires large-scale high quality face-text data.

Bidirectional Image-and-Text Generation. More recently, researchers have attempted to unify the text-to-image and image-to-text generation tasks in a single model. Huang et al. [18] formulate both the tasks as sequence generation tasks with a two-stage training strategy. ERNIE-ViLG [58] unifies pre-training method for the bidirectional image-text generation with language modeling objective and pre-trained a 10-billion parameter model on 145 million image-text pairs.

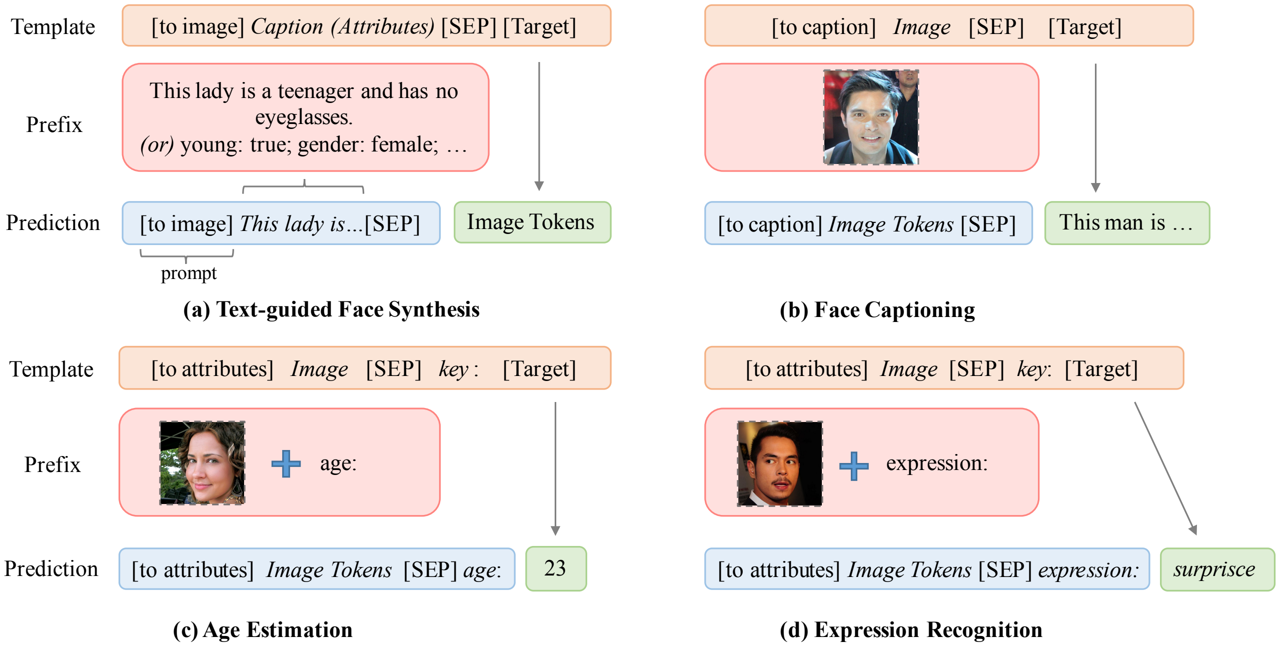
To the best of our knowledge, there is currently no bidirectional face-and-text generation work in face domain. The main reason lies in the complexity of facial attributes, and the difficulty of co-training with heterogeneous data from different tasks. In this work, we uniformly represent diverse face tasks with the sequence-to-sequence format, enabling knowledge sharing among different face tasks. Our work is the first one to provide a unified framework, for both text guided face synthesis and face captioning, and diverse facial analysis tasks like age estimation, expression recognition, race classification and facial attribute classification.

3 APPROACH

In this section, we introduce the architecture, input/output format and objective of our framework. Also, we present the details of data collection, prompt engineering and inference templates for applying the proposed framework to different face generation and analysis tasks. Figure 2 shows an overview of our model.

3.1 Unified Generative Framework

Text and Face Representation. For unifying input/output text and image, we represent both face and text as discrete sequences. Following previous NLP pre-training models [22, 24], the text is


Template
[to image] *Caption (Attributes)* [SEP] [Target]
Prefix
This lady is a teenager and has no eyeglasses.
(or) young: true; gender: female; …
Prediction
[to image] *This lady is*…[SEP]
Image Tokens
prompt
[to caption] *Image* [SEP] [Target]
[to caption] *Image Tokens* [SEP]
This man is …
Template
[to attributes] *Image* [SEP] *key* : [Target]
Prefix
age:
Prediction
[to attributes] *Image Tokens* [SEP] *age*:
23
[to attributes] *Image* [SEP] *key*: [Target]
expression:
[to attributes] *Image Tokens* [SEP] *expression:*
surprisce

(a) Text-guided Face Synthesis

(b) Face Captioning

(c) Age Estimation

(d) Expression Recognition

Figure 3: Inference templates for (a) text-guided face synthesis, (b) face captioning, (c) age estimation and (d) expression recognition.

tokenized to uncased subword tokens by WordPiece [50]. Similarly, the face image is "tokenized" to discrete visual tokens. The image tokenizer is the encoder of pre-trained discrete variational autoencoder (dVAE) [38] that encodes input image and quantizes it into discrete visual tokens according to a codebook (i.e., vocabulary). And the dVAE decoder is used to reconstruct the face image from visual tokens.

The model input is made up of three parts: prompt, prefix, and target. The prompt is always given first, and it informs the model what task to perform in words. The prompt is then followed by prefix and target i.e. the task context and the expected output. The three parts are concatenated into a training instance, connected by separate tokens (*[SEP]*). For each sample, we always add a special start-of-sequence token (*[SOP]*) at the beginning and a separate token at the end. The separate token marks the text (or image) boundary and is used to learn when to terminate the decoding process. The input sequence is divided into two segments: the prompt and prefix in the first segment, and the target in the second. Position embedding is used for retaining positional information. We use standard learnable 1D position embeddings for both text and face image [12, 22]. For each token in the input sequence, its vector representation is computed by summing the corresponding token, position, and segment embedding.

Training Objective. In our framework, all tasks are learned as a sequence-to-sequence task, that is, converting the source sequences from one domain (e.g., face image) to target sequences in another domain (e.g., caption, category).

The faces represented as sequences are treated as a "dialect" that shares the model structure and objective [36] with text. We use the transformer encoder for sequence modeling to predict tokens conditioned on the preceding tokens. Given a sequence of the concatenation of source tokens and target tokens, denoted as $\{t_1, ..., t_n\}$, we use the sequence modeling objective to maximize the following likelihood:

$$L = \sum_i \log P\left(t_i | t_1, ..., t_{i-1}; \theta\right) \tag{1}$$

where θ is parameters of a neural network. These parameters are trained using stochastic gradient descent.

Backbone Network. In our setting, we use a layer stacked transformer encoder [45] with a few modifications. Following Radford et al. [36], layer normalization layers are moved to the input of each sub-block. For each layer, the input vectors is linearly projected to a triple of queries, keys and values (Q, K, V). The output of a a self-attention head is computed via:

$$M = \begin{cases} 0, & allow\ to\ attend \\ -\infty, & prevent\ from\ attening \end{cases} \tag{2}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + M)\ V \tag{3}$$

where d_k is the dimension of keys. We use prefix attention mask M to control each token's access to its context [11]. The tokens in the first segment are allowed to bidirectionally attending other source tokens, and the second segment tokens are only allowed to attend to its left context.

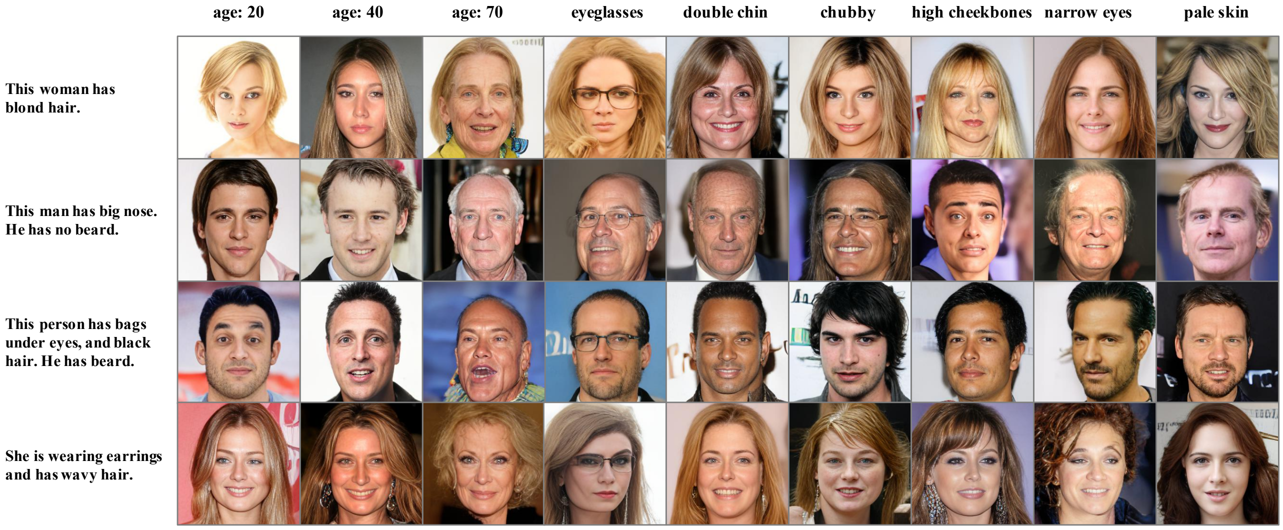

age: 20
age: 40
age: 70
eyeglasses
double chin
chubby
high cheekbones
narrow eyes
pale skin
This woman has blond hair.
This man has big nose. He has no beard.
This person has bags under eyes, and black hair. He has beard.
She is wearing earrings and has wavy hair.

Figure 4: Faces generated on different textual conditions. We use captions (left) and attributes (top) to jointly control diverse face generation.

3.2 General Cross-task Face-Text Paired Dataset and Talk2Face Training

We collect 13 datasets from various face tasks for learning generic face knowledge, as shown in Table 1. These datasets have different annotation formats and differ in granularity. For example, age labels are integers (e.g., *25*) in CACD, range (e.g., *20-30*) in Fairface, and text (e.g., *"This man is in his thirties."*) in CelebA-Dialog. In our work, all the annotations are converted into natural language without information loss.

For different modalities of supervision, we set up a protocol to reform them with natural language. For facial analysis tasks (e.g., classification and recognition), the annotated objects are converted into key-value pairs. The key is composed of [dataset name] and [attribute], which potentially identifies the value's data type, i.e., numbers or textual categories, or boolean. For example, *"cacd age: 25"* and *"fairface age: 10-20"*. Also, we observe that labels are sometimes defined differently in different datasets, e.g., the expression "surprise" has slightly different activation thresholds in ExpW [59] and AffectNet [32]. As a result, another purpose of the keys is to discern the inherent bias of overlapped datasets. Datasets with multiple attributes are joined by a semicolon ";", such as *"celeba gender: male; bald: true; oval face: false; ..."*.

We are able to straightforwardly cast all of the attributes we considered into key-value pairs with the exception of FS2K [14], where eyes color and lips color are RGB values like *"188,143,142"*. We use the RGB Color Codes Chart to find the closest color name, for example, *"188,143,142"* → *"rosy brown"*.

For text-and-image datasets, the image relating to multiple texts is separated into many one-to-one image-text pairs. In this way, different face datasets are unified with matched face and text pairs to form a face-generic dataset. In total, our dataset contains 2.3 million face-text pairs collected from the aforementioned datasets.

Table 1: Face Datasets

Dataset	Samples	Supervision
CelebA [29]	203k	40 facial attribute annotations
FS2k [14]	2k	24 features depict diverse scenes
AffectNet [32]	400k	8 facial expressions
RAF-DB [27]	29k	7 emotions
ExpW [59]	92k	7 facial expressions
CACD [7]	163k	age of 2,000 celebrities
AFAD [33]	160K	age and gender of Asian
IMDB-WIKI [41]	523k	age and gender
FairFace [20]	101k	race, gender, and age
FFHQ-Text [62]	8k	text description (manually)
MM-CelebA-HQ [51]	300k	text description (38 attributes)
CelebAText-HQ [44]	150k	text description (manually)
CelebA-Dialog [19]	203k	text description (5 attributes)

In the training phase, the instance is made up of a prefix and target randomly sampled from our dataset, and a corresponding prompt. The prompt is mainly used to indicate the modality of the target (i.e., *[to image]*, *[to caption]* and *[to attributes]*). We don't specify the prefix because we expect the model to figure out automatically. Therefore, the training contains of a total of four tasks, namely face-to-caption, caption-to-face, face-to-attributes and attributes-to-face. These tasks are all represented in sequence-to-sequence format, which allows us to evaluate gradients and perform parameter update using the same training objective. During training, each minibatch has balanced training instances for each task.

3.3 Uniform Inference Template

For inference, Talk2Face tackles all tasks in a generative manner, i.e., directly generating text/face as output. We feed the model with prompt and prefix, then autoregressively sample tokens from model likelihood. This can be achieved by simply taking the token with

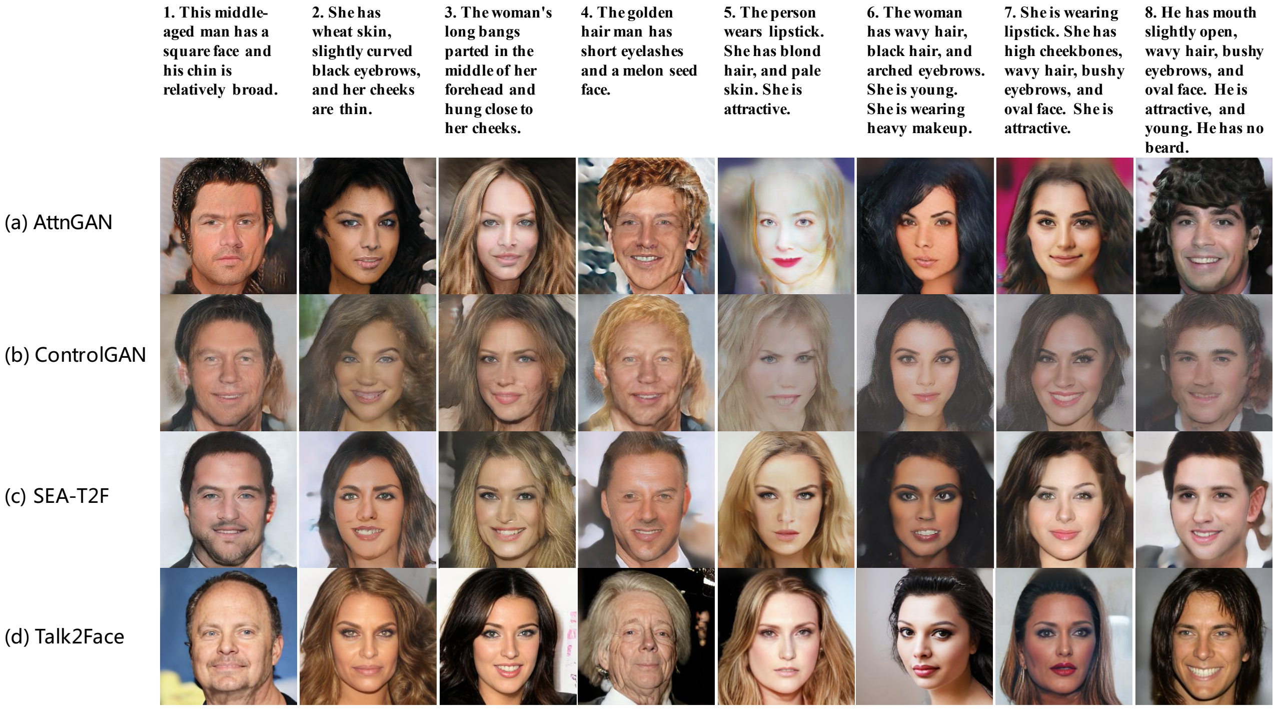

1. This middle-aged man has a square face and his chin is relatively broad.
2. She has wheat skin, slightly curved black eyebrows, and her cheeks are thin.
3. The woman's long bangs parted in the middle of her forehead and hung close to her cheeks.
4. The golden hair man has short eyelashes and a melon seed face.
5. The person wears lipstick. She has blond hair, and pale skin. She is attractive.
6. The woman has wavy hair, black hair, and arched eyebrows. She is young. She is wearing heavy makeup.
7. She is wearing lipstick. She has high cheekbones, wavy hair, bushy eyebrows, and oval face. She is attractive.
8. He has mouth slightly open, wavy hair, bushy eyebrows, and oval face. He is attractive, and young. He has no beard.
(a) AttnGAN
(b) ControlGAN
(c) SEA-T2F
(d) Talk2Face

Figure 5: Qualitative comparison of text-guided face synthesis.

the largest probability (greedy search) or using other stochastic sampling techniques. The sampling ends when the *[SEP]* token is generated. After that, we detokenize the generated sequence into an face image or text.

The model learns face-domain generic knowledge in the training stage. As a result, we can leverage knowledge from the model without fine-tuning when performing downstream tasks. For each task, we design an inference template that closely match the training task. The users only need to choose a template, fill in the context, and then let the model generate the *[Target]* part as the output.

Text-guided Face Synthesis. Figure 3-(a) presents a template for text-guided face synthesis task, where the textual condition can be caption or attributes or a combination of both, e.g., *"This man has black hair. smiling : true ;"*. Guided by the prompt *[to image]* and prefix, the model predicts a fixed-length sequence of image tokens, and then detokenized into a face image.

Face Image Captioning task takes an image as the prefix; the model generates a sequence of text tokens conditioned on the prompt *[to caption]*. Text tokens are then mapped to subwords and post-processed to a sentence. The inference template is shown in Figure 3-(b).

Facial Analysis. In Figure 3-(c) we present an example for age estimation. If the model outputs a string corresponding to a number at inference time, we convert it to an integer value; otherwise, we treat the model's prediction as incorrect. For expression classification (Figure 3-(d)), the model outputs a word representing the expression category. If the word is not in the candidates, it is considered incorrect.

We adopt different decoding strategies for autoregressive sampling depending on whether the labels of a given task are enumerable or not. For open-answer tasks such as face synthesis and face captioning, we adopt top-k/top-p filtering with temperature [16] to enrich the diversity of generated token sequences. We use the CLIP score [35] to select the top n outputs closest to the condition. For multiple-choice tasks like facial analysis, we anticipate the model to generate values inside the candidates, where accuracy takes precedence over diversity. Therefore, we obtain the results using greedy decoding (i.e., choosing the highest-probability token at every timestep).

The inference process is consistent for both text and image generation across different tasks, allowing downstream tasks to share the same sequence modeling head without extra parameters. Moreover, our model can be easily extended to new tasks by designing the corresponding templates with same rules.

4 EXPERIMENTS

We first train the Talk2Face model on our collected dataset and then evaluate it on different downstream tasks. We conduct experiments on face tasks including text guided face synthesis, face captioning and facial analysis e.g., age estimation, expression recognition, race categorization and face attribute classification.

4.1 Model Settings

We implement Talk2Fase based on UER-py framework [60] [1].

[1] https://github.com/dbiir/UER-py

Table 2: Quantitative comparison of text-guided face synthesis. ↓ means the lower the better while while ↑ means the opposite. Qual. and Cons. are abbreviations for quality and consistency.

Method	Metrics		User Study	
	FID ↓	IS ↑	Qual. ↑	Cons. ↑
AttnGAN [53]	126.0	**2.7**	3.4	3.1
ControlGAN [26]	116.3	1.9	3.7	3.7
TediGAN [51]	106.4	-	4.0	4.1
SEA-T2F [44]	-	1.9	3.8	3.6
Talk2Face (Ours)	**104.9**	2.2	**4.1**	**4.3**

Tokenizer. Our image tokenizer follows VQGAN [13], which has been pre-trained on FFHQ [21] dataset to better adapt to the distribution of faces. Each image is resized to 256 × 256 resolution and tokenized into 256 visual tokens with f=16 (frame size) and V=1024 (vocabulary size). The text tokenizer directly follows the BERT [22] model, which contains 30,522 uncased tokens. In total, our multimodal vocabulary contains 31,546 tokens.

Configuration. We use a 12-layer transformer with 768 hidden size, and 12 attention heads. The intermediate size of feed-forward networks is 3072. This results in a model with about 110 million parameters, which is comparable to the BERT-base. For regularization, we use a dropout with 0.1 probability.

Hyperparameters. The training runs for 500,000 steps with a batch size of 192. AdamW [30] with $\beta 1 = 0.9$, $\beta 2 = 0.999$ is employed for optimization. The learning rate is set to $2e^{-4}$ with linear warmup. The training process take about 15 days using 8 Nvidia Tesla P40 24GB GPUs.

4.2 Text-guided Face Synthesis

We first evaluate our model for text-guided face synthesis. The textual captions and attribute labels or a combination of both can be used as the control signals, opening up many possibilities for face synthesis.

In Figure 4, we show the synthesized results by simultaneously using textual captions and attribute labels. As can be observed, our method can generate photo-realistic and text-relevant faces. Moreover, our model can naturally combine textual captions and facial attributes for more controllable face synthesis.

Quantitative Evaluation. We adopt Fréchet Inception Distance (FID) [15] and Inception Score (IS) [42] to measure the statistical similarity between the generated faces and real ones. Specifically, we calculate FID and IS scores between 1,000 generated faces and the validation split of our dataset. We compare our method with existing text-to-face generation approaches including AttnGAN [53], ControlGAN [26], TediGAN [51] and SEA-T2F [44]. The results are shown in Table 2. We can see that our method outperforms other approaches for both metrics.

In addition, we also conduct a user study to manually evaluate the image quality and consistency with the given textual captions. In our experiment, we show each user a given textual caption and a corresponding face generated by different approaches, and ask them to rate a score between 1-5 in terms of the image quality and consistency with the given text. We collect 50 text-face pairs and invite 10 users to perform this task. The average results are shown in Table 2, there is an obvious preference for our approach over others.

Qualitative Comparison. The visual comparisons of different methods on Multi-Modal CelebA-HQ [51] and CelebAText-HQ [44] are shown in Figure 5. Row (a)-(d) present the example faces generated by four methods, and column 1-8 represent the text on which the images are generated.

We find faces generated by AttnGAN, ControlGAN with blurry textures and color distortions. Look into the details of SEA-T2F, where some face attributes are lost or blurred (e.g., ears of **c1** and teeth of **c2**), some appearances are not photo-realistic (e.g., hair of **c5** and face shape of **c1**). In addition, we also find that existing approaches have difficulty in synthesizing regions around the face, including the neck and clothes, (e.g., neck of **a5**, **b5**, and **c8** are blurred, and the right collar of **c1** is missing). In comparison with other approaches, Talk2Face generates clearer and more photo-realistic faces.

We further evaluate the synthesis diversity of our method. In particular, we compare multiple synthesized face images using the same text description with the state-of-the-art GAN-based method TediGAN [51]. As shown in Figure 6, we can see that the samples generated by TediGAN are very similar, except for the color of skin and background. By contrast, the proposed Talk2Face can produce more diverse faces with different ages and ethnicities. On the other hand, in order to generate different samples, TediGAN requires style mixing with manually selected layers. By contrast, the diversity of our method comes from a random sampling strategy without human interaction.

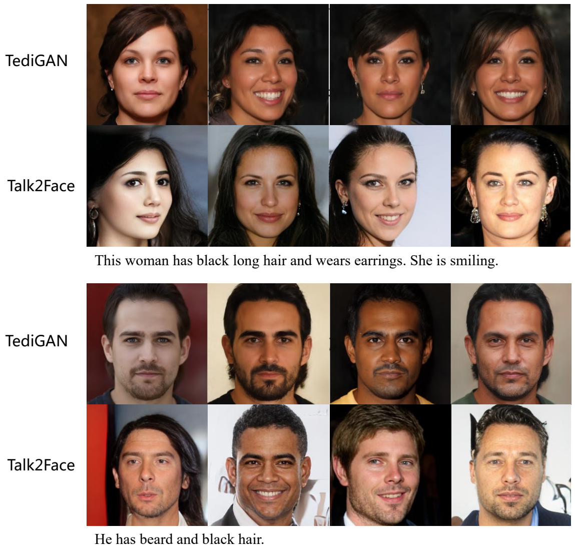
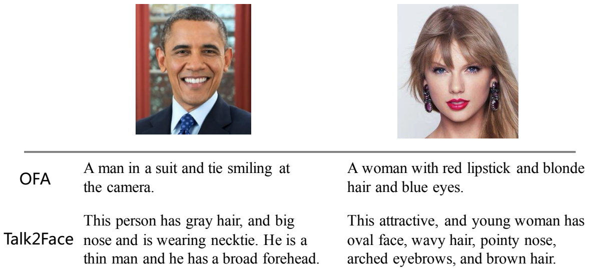

TediGAN
Talk2Face
This woman has black long hair and wears earrings. She is smiling.
TediGAN
Talk2Face
He has beard and black hair.

Figure 6: Diversity comparison with TediGAN.

4.3 Face Image Captioning

In this part, we evaluate the proposed model for face image captioning. We compare Talk2Face with recent state-of-the-art method



OFA	A man in a suit and tie smiling at the camera.	A woman with red lipstick and blonde hair and blue eyes.
Talk2Face	This person has gray hair, and big nose and is wearing necktie. He is a thin man and he has a broad forehead.	This attractive, and young woman has oval face, wavy hair, pointy nose, arched eyebrows, and brown hair.

Figure 7: Generated captions on face images. Compared to OFA, Talk2Face can generate more detailed captions.

Table 3: Results of face captioning on CelebA-text. B-4 and R-L are abbreviations for BLEU-4gram and ROUGE-L.

Method	B-4	R-L	METEOR	CIDEr	Rich.	Acc.(%)
OFA	12.5	33.6	15.9	11.4	2.9	87.2
Talk2Face (Ours)	**33.4**	**53.5**	**28.4**	**40.0**	**3.8**	**92.7**

Table 4: Performance on facial analysis tasks. Note that baselines are modeled for single task, whereas our model generalizes to all tasks. AE: age estimation, ER: expression recognition, RC: race classification, MAC: multi-attribute classification.

Models	AE	ER	RC	MAC
	MAE ↓	Acc. ↑	Acc. ↑	Acc. ↑
CORAL[4]	**5.4**	-	-	-
Mollahosseini et al. [32]	-	**58.0**	-	-
Karkkainen and Joo [20]	-	-	**75.4**	-
HFE[55]	-	-	-	**92.2**
Talk2Face (Ours)	8.3	44.7	66.2	90.3

OFA [48] on CelebAText-HQ [44] test set, which is manually annotated and therefore closer to the general image caption task. We use BLEU[34], ROUGE[28], CIDEr [46], METEOR[1] as evaluation metrics.

Human evaluation is also conducted to assess the richness and accuracy of generated captions. The richness is determined by the number of entities (attributes) stated in the caption, and the accuracy is assessed by the entities correctly described. We randomly select 50 images to generate captions and ask 5 users to annotate entities and accuracy. As illustrated in Table 3, Talk2Face outperforms the baseline model on all metrics. Qualitative examples are provided in Figure 7. It can be seen that the captions generated by Talk2Face are much richer and more accurate.

4.4 Facial Analysis

We now evaluate Talk2Face on four facial analysis tasks, i.e., age estimation on Cross-Age Celebrity Dataset (CACD) [7], expression recognition on AffectNet [32], race classification on FairFace [20] and multi-attribute classification on CelebFaces Attributes dataset (CelebA) [29].

We evaluate age estimation with mean absolute error (MAE), and accuracy for other tasks. In Table 4, we compare Talk2Face with approaches specially designed and trained for different tasks. As shown in the table, while the multiple attribute classification accuracy of Talk2Face is comparable to that of HFE, the performances of our approach on AE, ER and RC are not as good as that of STOA methods.

One of the possible reasons is that our method needs to simultaneously adapt to different classification labels of multiple datasets. For example, in the expression recognition task, the number of expressions labeled for AffectNet and ExpW are 8 and 7, respectively. As a result, it is more difficult to learn a unified model than a task-specified one. The second reason is the capability of the BERT tokenizer in numerical representation. The BERT vocabulary represents each integer within 200 as separate word, therefore, in the age estimation task, it's difficult for the loss function to measure the numerical differences between the label and the prediction. One possible improvements is to modify the text tokenizer to map each digit to a word (e.g., "19" is tokenized into "1" and "9").

The reason for better performance of Talk2Face on CelebA could be that, CelebA transforms multi-label classification into multiple binary-classification tasks, allowing the model to focus on a relatively simple subtask.

In this study, we do not aim to boost performance of our model on a specific task; instead, we seek a possibility of a unified model for diverse face generation and analysis tasks. We hope our work could inspire future works in this direction.

5 CONCLUSION

In this paper, we present a general generative framework, Talk2Face, for a number of different face generation and analysis tasks, e.g. text-guided face synthesis, face captioning, age estimation, expression recognition and attribute classification etc. Based on the unified sequence representation for text, face and category labels, the model trained with 2.3 million face-text pairs achieves comparable performances with those models specially designed and trained for specific tasks, without any fine-tuning.

Limitations and future work. The current limitation of our approach is that creating faces in an autoregressive manner is computationally expensive because images are two-dimensional and have considerably more tokens than text. We aim to use linear time transformer variants to speed up training and inference of long sequences in future work. Although our model is versatile, we do not currently support face editing tasks (text + image ⟶ image in sequence-to-sequence format). In the future, we will scale up the training to support more downstream tasks.

6 ACKNOWLEDGMENTS

This research was supported by National Natural Science Foundation of China under grant no. 91959108 and Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515111199 and 2022A1515011018.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings*

of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 65–72.

[2] Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters* 140 (2020), 325–331.

[5] Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, and Yuval Elovici. 2022. Fair and accurate age prediction using distribution aware data curation and augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3551–3561.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* 33 (2020), 9912–9924.

[7] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. 2014. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*. Springer, 768–783.

[8] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852* (2021).

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021).

[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems* 32 (2019).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.

[14] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. 2021. Deep Facial Synthesis: A New Challenge. *arXiv preprint arXiv:2112.15439* (2021).

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[16] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

[17] Xianxu Hou, Xiaokang Zhang, Yudong Li, and Linlin Shen. 2022. TextFace: Text-to-Style Mapping based Face Generation and Manipulation. *IEEE Transactions on Multimedia* (2022).

[18] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. 2021. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1138–1147.

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-Edit: Fine-Grained Facial Editing via Dialog. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[20] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.

[21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[23] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. 2021. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790* (2021).

[24] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems* 32 (2019).

[27] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2584–2593.

[28] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[30] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[31] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition. In *Proceedings of the International Conference on Artificial Intelligence (IJCAI)*.

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[33] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4920–4928.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[39] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. 2017. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 17–24.

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[41] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*. 10–15.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2234–2242.

[43] Andrey V Savchenko. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 119–124.

[44] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. 2021. Multi-caption Text-to-Face Synthesis: Dataset and Algorithm. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2290–2298.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[46] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[48] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv preprint arXiv:2202.03052* (2022).

[49] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. 2021. Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition. *arXiv preprint arXiv:2109.07270* (2021).

[50] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[51] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards Open-World Text-Guided Face Image Generation and Manipulation. *arxiv preprint arxiv: 2104.08910* (2021).

[53] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.

[54] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2579–2591.

[55] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. 2020. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13055–13064.

[56] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1947–1962.

[58] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViLG: Unified Generative Pre-training for Bidirectional Vision-Language Generation. *arXiv preprint arXiv:2112.15283* (2021).

[59] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision* 126, 5 (2018), 550–569.

[60] Zhe Zhao, Hui Chen, Jinbin Zhang, Wayne Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An Open-Source Toolkit for Pre-training Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. 241–246.

[61] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.

[62] Yutong Zhou and Nobutaka Shimada. 2021. Generative Adversarial Network for Text-to-Face Synthesis and Manipulation with Pretrained BERT Model. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 01–08.

Table 5: Effect of components in Talk2Face.

Models	CelebA-HQ FID ↓	CACD MAE ↓	CelebA Acc. ↑
Talk2Face (100k steps)	**106.2**	10.9	87.2
– OpenImage VQGAN	141.7	**10.3**	**88.4**
– Single-dataset (CACD)	-	11.7	-
– Single-dataset (CelebA)	-	-	86.2

A ANALYSIS

In this part, we analyse the effect of image tokenizer and multi-dataset training. The models are evaluated on face synthesis (i.e., CelebA-HQ), age estimation (i.e., CACD) and multi-attribute classification (i.e., CelebA). We set the training steps to 100,000, which is 20% of the total steps used in the previous experiments. Table 5 reports the results.

Effect of Image Tokenizer. We replace Talk2Face's image tokenizer with VQGAN-OpenImage (f=8, V=8192), which is pretrained on a large-scale dataset beyond face images. Thus, it has a large codebook and sequence length. We find that it performs better on face analysis tasks, but not so well on face synthesis. The main reason is that longer input sequences can provide more detailed image representations that are helpful for classification tasks, but are more challenging to produce as outputs. Qualitative examples of different VQGAN are shown in Figure 8.

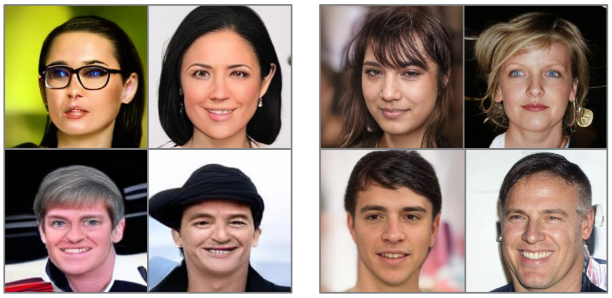

(a) VQGAN-OpenImage (b) VQGAN-FFHQ

Figure 8: Qualitative examples of different image tokenizers. (a) Samples form VQGAN-OpenImage. (b) Samples form VQGAN-FFHQ.

Single-dataset vs. Multi-dataset Training. We compare Talk2Face trained on a single dataset and our large-scale dataset. For single-dataset models, we keep the preprocessing protocol unchanged and separately use CACD and CelebA data for training. It can be seen that multi-dataset training outperforms that of single-dataset. We assume that heterogeneous datasets provide generic knowledge that is beneficial for learning specific tasks.

B SELF-ATTENTION MAP

Multi-head self-attention allows the model to jointly attend to information from different representation subspaces at different positions. In Talk2Face, we observe that each attention head focuses on different regions of the image and is able to distinguish between the person and background.

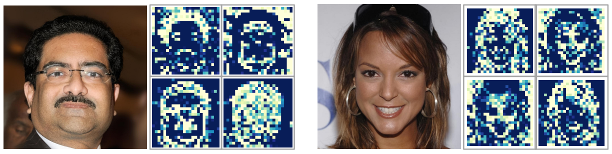


Figure 9: Self-attention map for different attention heads.

In Figure 9, we plot the self-attention map of different heads within an image. The visualizations are produced by attention scores computed via query-key product in the last layer at the last time-step. It shows the image tokens that Talk2Face attends to when predicting the next token. Talk2Face learns to distinguish semantic regions from text-image pairs, even if there are no image segmentation annotations in the training data. Similar properties are also observed by Caron et al. [6] and Bao et al. [2]. It partially indicates Talk2Face has learned generic face knowledge.