ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Hi Sheldon! Creating Deep Personalized Characters from TV Shows

Anonymous ICCV submission

Paper ID 3137

Abstract

Imagine an interesting multimodal interactive scenario that you can see, hear, and chat with an AI-generated digital character, who is capable of behaving like Sheldon from The Big Bang Theory, as a DEEP copy from appearance to personality. Towards this fantastic multimodal chatting scenario, we propose a novel task, named Deep Personalized Character Creation (DPCC): creating multimodal chat personalized characters from multimodal data such as TV shows. Specifically, given a single- or multi-modality input (text, audio, video), the goal of DPCC is to generate a multi-modality (text, audio, video) response, which should be well-matched the personality of a specific character such as Sheldon, and of high quality as well. To support this novel task, we further collect a character centric multimodal dialogue dataset, named Deep Personalized Character Dataset (DPCD), from TV shows. DPCD contains character-specific multimodal dialogue data of 10k utterances and 6 hours of audio/video per character, which is around 10 times larger compared to existing related datasets. On DPCD, we present a baseline method for the DPCC task and create 5 Deep personalized digital Characters (DeepCharacters) from Big Bang TV Shows. We conduct both subjective and objective experiments to evaluate the multimodal response from DeepCharacters in terms of characterization and quality. The results demonstrates that, on our collected DPCD dataset, the proposed baseline can create personalized digital characters for generating multimodal response. Our collected DPCD dataset, the code of data collection and our baseline will be published soon.

1. Introduction

Advancements in artificial intelligence have recently enabled highly realistic content synthesis, such as text to speech [32, 2, 26], text to image [30, 33, 34, 29], text to 3D [25], text to video [36], speech-driven facial animation [43, 19, 40, 35], etc. These AI techniques lay the foundation of AI-generated characters, which have wide application in social services such as education, healthcare, and
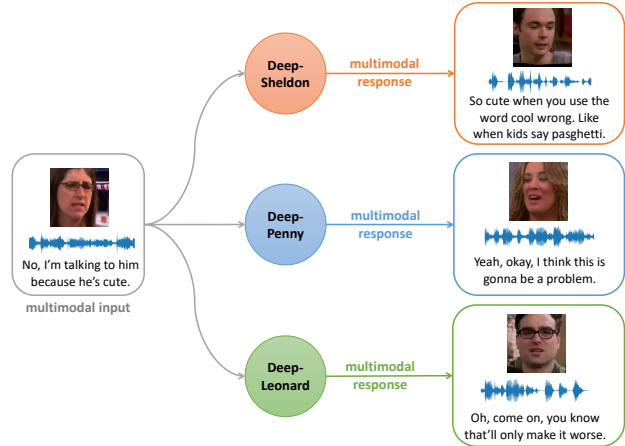

Figure 1. Our proposed Deep Personalized Character Creation (DPCC) task aims to create DeepCharacters, providing charaterized multimodal response, given the same stimuli.

so on [24]. We expect to interact with an AI-generated personalized character, just feels like interacting with a specific person, and the AI-generated character can respond to you with personalized text, voice and facial expression. Instead of creating a totally factitious character from manual settings, in this paper, we are heading for a novel setting: creating a *DEEP* personalized specific character from his/her dialog videos providing textual, vocal and visual data. By *DEEP*, we mean not only mimicing the *outer* characteristics such as appearance and tune, but also the *inner* characteristics such as personality. Once we get a *DEEP* personalized Sheldon (*DeepSheldon*), one can have a multimodal chat with him and feel like we are really talking to the one you know from The Big Bang Theory. This technique of creating this kind of specific personalized digital characters has commercial demand in entertainment industry, such as generating virtual content of characters in TV shows, movies, games and concerts. One can imagine an interesting and exciting scenario that the AI-generated characters interact with each other and continue their story in a visual world line.

Towards this fantastic goal, we propose a novel task, named Deep Personalized Character Creation (DPCC), as

ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

generating a personalized character from the multimodal conversation data including text, audio, and video. As shown in Figure 1, DPCC takes the multimodal stimuli as input, and responds with a personalized multimodal output. Given the same input, different Deep personalized digital Characters (DeepCharacters) should provide responses that well reflect their individual trait, respectively. With multimodal response, DeepCharacters can bring more concrete and immersive interactive experience, compared to previous chatbots, which can only provide text or voice response. DPCC can also support text-only input and generate multimodal response. Please note that this task is different from text or voice-driven generation, which is mainly converting from one modality to another [43, 19, 40, 35]. In our DPCC, the input is the stimuli, no explicit content of the response is provided, and each modality content of the response needs to be predicted. Also, different from previous personalized conversation generation tasks, our DPCC doesn't require additional profiles or traits as input to explicitly label personality, and learn the character's interaction pattern, speaking style and personality from conversation contexts.

To support DPCC task, we collect a multimodal conversation dataset, named Deep Personalized Character Dataset (DPCD). DPCD containing 5 main characters from *The Big Bang Theory*[1], and in total 29.75 hours videos and 29k conversation turns. We only collect the video clips containing characters' talking face to extract effective facial expression features. Besides, we manually check the auto-alignment between text and audio/video modality. We also conduct preprocessing to remove noise in audio. Those efforts ensure the collected data is of high quality for DPCC task. There do exist multimodal (text, audio, video) datasets [4, 27, 10, 11, 5] for natural language processing (NLP), audio-visual synchronization (AVS) and speech synthesis. These existing multimodal datasets are collected for a general understanding of conversation content, emotion, or general synthesis of speech and talking head that can be applied to different persons, not for modelling specific characters. Consequently, the data for modeling the personality of a specific character is typically insufficient. For some datasets, the character labels of conversations are even not provided. As shown in Table 1 in Section 3.2, the average number of utterances for each character is at most 1,000 (on IEMOCAP [4]), while our DPCD focus on the main characters and can provide almost an order more (9,718 utterances per character). Besides, those datasets are often of low quality in terms of text/audio/video misalignment, noisy audio, and face missing.

We present a baseline solution for DPCC task and create 5 DeepCharacters from the collected DPCD. To evaluate the textual and multimodal response from DeepCharacters in terms of characterization and quality, we design some tools to evaluate the characterization quality. Both subjective and objective experimental results demonstrate that the response of our created DeepCharacters can well reflect the personality of characters, indicating DPCD can be used for DPCC task.

Our main contributions can be summarized as: (i) We introduce a novel tasks to create deep personalized specific characters, who can generate the personalized multimodal response. We believe this task can facilitate the research on multimodal and personalized AI generation. (ii) We collected a new dataset for the proposed task, which fills the gap of lacking high-quality, large-scale and character-centric conversation data. We think this dataset can benefit the community and step further towards personalized character creation. (iii) We provide a baseline solution for this proposed task, and design some tools to evaluate the created DeepCharacters in terms of both characterization and quality. This lays the foundation of how to create and evaluate DeepCharacters for future works.

2. Related Work

Modeling personalized conversation has long been a complicated issue. Li et al.[18] address the challenge of persona consistency by encoding personas and adding persona embeddings into seq2seq model. They generate persona-specific responses based on the conversations of 13 characters from TV series *Friends* and *The Big Bang Theory*. Based on the persona-based model by Li et al.[18], Xing et al.[45] estimate the OCEAN scores for each speaker to train persona embedding.

Considering the difficulty and complexity of persona-based conversation modeling, many models rely on extra profile or personality traits to develop a sense of persona. Mazaré et al.[21] define persona as a set of sentences representing the personality of the responding agent and train persona-based dialogue models on a large-scale persona-based dialogue dataset (PCR) extracted from REDDIT[2]. Zhong et al.[53] further extend PCR with annotated empathy information. Zheng et al.[52] build a persona-based empathetic conversation response selection model from amounts of multi-turn empathetic conversation data with five personality annotations (i.e.Gender, Age, Location, Interest Tags, and Self-description) for each speaker. Wu et al.[44] propose a generative split memory network to incorporate diverse personal information and generate personalized responses based on a personalized dataset PER-CHAT with finer-grained user personal information and contextual comments. Chen et al. [7] build the personality detection and emotion recognition model from a Chinese personalized dataset CPED with annotations of speakers' personalities, dynamic emotions and dialog actions.

---

[1]https://bigbangtrans.wordpress.com/

[2]https://www.reddit.com/r/datasets/ comments/3bxlg7/

Constrained by the lack of single speaker's conversation data and the complexity of personality modeling, the models listed above mainly focus on personal information consistency rather than character consistency, and rely heavily on detailed profiles. Different from them, we directly train a deep personalized character model from a rather larger volume of multimodal conversation data for a single speaker. Detailed comparisons of our multimodal dataset DPCD to the related datasets are presented in Table 1 in Section 3.2.

3. Deep Personalized Conversation Dataset

3.1. Overview of DPCD Collection

Our goal is to collect character-centric multimodal conversations, from raw videos, audio, transcripts, and subtitles from the TV show. We choose *The Big Bang Theory* as a example in this paper, for its large number of episodes. The collected items should well reflect the characters' clean contents including facial expressions, vocals and texts. Meanwhile, this content should be paired and temporally well aligned to allow cross-modality integration. To conduct such a large scale data collection and ensure high quality, we adopt an automatic and manual mixed operation as follows: 1) For each modality, we use automatic tools to extract the basic feature, reduce the noise, and roughly align modalities. 2) Manually check the quality of the alignment between text and audio/video modality. During the automatic processing, for videos, we recognize and track the speaking face, then verify the active speaker and annotate its identity. For audio, we use multiple filters and acoustic tool kits to reduce background noise and enhance vocals, and we also remove laughter from the audience to get clean acoustic signals. As for the text, we clean and align the transcripts and subtitles through textual similarity computation. For feature extraction, we use OpenFace[3] to extract facial expression features, COVAREP[4] to extract acoustic features, and DialoGPT tokenizer[51] to get word embeddings. Finally, we align all three modalities according to the starting and ending timestamps. Thanks to the manual efforts, the final dataset includes well-aligned visual, audio, and textual modalities for each utterance, as well as the speaker identity and conversational contexts. Details of the data collection are presented in Supplement Section 1.1 and 1.2. Please note that this collection pipeline can be applied to other TV shows or movies to scale the dataset, and we have verified the effectiveness of the pipeline on *Friends*, see Supplement Section 1.3 for details.

3.2. Dataset Exploration

We collect single utterances that are perfectly aligned with corresponding audios to train speech synthesis models.

---

[3]https://github.com/cmusatyalab/openface
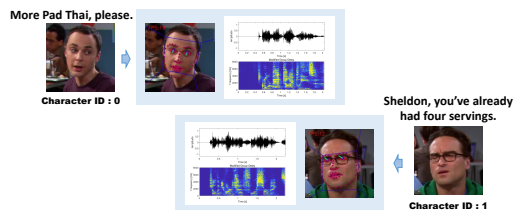[4]https://www.adobe.com/cn/products/premiere.html

Figure 2. An example from our dataset. A sentence in a dialogue is aligned with corresponding audio, video, speaker identity and extracted visual/acoustic features.

Table 2 presents several key statistics of the DPCD dataset. We collect 48,589 utterances from 5 characters (Sheldon, Leonard, Penny, Howard, Raj), each utterance contains 7.06 words and lasts 2.20s on average. Since we divide sentences on the basis of subtitles, the average duration for an utterance is relatively short. However, the total duration of audio is 29.65h, reaching a volume qualified for speech synthesis task[32, 31, 6].

We also maintain complete conversational order of the utterances, and extract mono conversation turns as an example. Using these conversation data, we can learn character's interaction pattern and speaking style, and try to model that character's personality. Figure 2 gives a brief example of conversation data in our DPCD dataset. Our DPCD dataset collects 29,235 mono conversation turns in total. Each conversation turn contains about 27.38 words and lasts 10.17s. As conversations involve interaction between speakers, the corpus has a richer collection of utterances and is more expressive.

Detailed comparison of DPCD to the related datasets is presented in Table 1. According to the targeting tasks, we roughly divide the multimodal datasets into three categories for natural language processing (NLP), audio-visual synchronization (AVS) and speech synthesis, respectively. The existing multimodal datasets for NLP [49, 3, 47, 12, 4, 27, 10, 11, 5, 20] are usually designed to understand interaction patterns or detect the fine-grained conversation features in dialogue turns. However, they are typically collected from a wide range of speakers to balance gender and race, resulting in inadequate data per speaker for tasks like personalized character encoding. Audio-visual datasets [22, 41, 8] for AVS typically consist of huge amounts of high-quality talking head videos. However, since it's rather challenging to extract subtitles from in-the-wild videos, these datasets hardly contain aligned texts or conversation contexts, and thus are not qualified for most conversation tasks. Datasets for speech synthesis [14, 46] usually contain elaborately aligned text and acoustic modalities. They are always recorded in a noiseless studio to guarantee acoustic quality. Transcripts are forced aligned with audio clips to maintain accurate correspondence relationship between phonemes and pronunciation. All of the existing multi-

ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Dataset	Modalities			Extra Attributes			Data Source	# of utterances	Total duration	# of speakers	Avg.# of utterances per speaker	Avg.dur.of utterances per speaker	Qualified Tasks			
	V	A	T	Conversational	E/S Label	Character Tag							NLP	AVS	Speech Synthesis	DPCC
CMU-MOSI [49]	✓	✓	✓		✓		YouTube	2199	2.56h	89	24.71	0.029h	✓			
CMU-MOSEI [3]	✓	✓	✓		✓		YouTube	23453	65.89h	1000	23.45	0.066h	✓			
IEMOCAP [4]	✓	✓	✓	✓	✓		Role Play	10000	11.47h	10	1000	1.47h	✓			
MELD [27]	✓	✓	✓	✓	✓		TV Series	13708	12.68h	407	33.68	0.031h	✓			
MEISD [10]	✓	✓	✓	✓	✓		TV Series	20000	22.22h	4072	4.91	5.46e-03h	✓			
SEMD [11]	✓	✓	✓	✓	✓		TV Series	1,066,677	746h	-	-	-	✓			
MSCTD [20]	✓			✓	✓		Former Dataset	173,241	-	-	-	-	✓			
MUStARD [5]	✓	✓	✓		✓	✓	YouTube,MELD and TV Series	14351	20.81h	23	623.96	0.9h	✓			
UR-FUNNY [12]	✓	✓	✓		✓		TED Talks	63727	90.23h	1741	36.6	0.052h	✓			
CH-SIMS [47]	✓	✓	✓		✓		TV Series	2281	2.33h	474	4.81	4.92e-03h	✓			
CPED[7]	✓	✓	✓	✓	✓	✓	TV Series	132,762	78.34h	392	338.68	0.20h	✓			
Voxceleb1 [22]	✓	✓				✓	YouTube	145,116	352h	1251	116	0.28h		✓		
Voxceleb2 [22]	✓	✓				✓	YouTube	**1,130,720**	2442h	**6112**	185	0.40h		✓		
MEAD [41]	✓	✓					Role Play	-	38.95h	60	-	0.65h		✓		
LRW [8]	✓	✓	✓				TV Programs	-	**1000+h**	1000+	-	-		✓		
LJ-Speech [14]		✓	✓			✓	Book Recording	-	24h	1	-	-			✓	
CSTR VCTK [46]		✓	✓			✓	Book Recording	-	44h	110	-	-			✓	
DPCD (Ours)	✓	✓	✓	✓		✓	TV Series	48589	29.65h	5	**9717.8**	**5.93h**	✓	✓	✓	✓

Table 1. A detailed comparison between our dataset DPCD and other multimodal datasets.(E/S Label: Emotion/Sentiment Label)

Aligned Utterances (WV)		Conversations (W5-R)	
# of utterance	48589	# of mono conversations	29235
# of speaker	5	# of words	800356
dura of audios(h)	29.65	Avg. # of words per turn	27.38
Avg. words per utterance	7.06	Avg. dura of mono turn(s)	10.17
Avg. # of utterance per speaker	9717.8	Avg. # of utterances per mono turn	3.67
Avg dura of audio per speaker(h)	5.93	Avg. # of conversation per speaker	5847

Table 2. Perfectly aligned utterances with videos (WV) and all conversations with 5 main characters as respondent (W5-R).

Detailed Statistics	Sheldon	Penny	Leonard	Howard	Raj
# of utterance	18747	7636	9803	6557	5846
dura of audios(h)	12.10	5.68	4.30	4.07	3.49
Avg. words per utterance	7.05	7.09	7.05	7.05	7.13
# of conversation turns	8564	5566	7075	4434	3633
Avg. words per response	17.32	12.05	12.01	13.01	13.84

Table 3. Detailed statistics on characters in DPCD.

modal datasets are always collected to qualify a specific task in a single field, hence leaving some modalities of poor quality. With the total scale of data increasing, the data size for each character is often ignored.

Different from most multimodal datasets targeting NLP tasks, the videos and audios offered by our DPCD dataset are more informative and clean. Our videos are specifically cropped to focus on active talking face, excluding other characters and complicated background scenes. So the visual features extracted from these video crops can better represent facial expressions and the speaker's emotions, free from possible environmental noises. The audios in our dataset are also elaborately filtered to better capture the speaker's voice and tone. Compared with many audio-visual datasets, our texts are labeled with speaker identity and conversational context, available for further expression and emotion modeling. In this way, our dataset DPCD is qualified for tasks in the fields such as natural language processing, audio-visual synchronization and speech synthesis.

Furthermore, as mentioned above, our DPCD dataset mainly focuses on 5 characters, detailed statistics for each character are presented in Table 3. It is obvious, from the table, that our dataset offers a competitively large volume of data for a single character. Sheldon acted as a respondent in 8,564 mono conversation turns, each containing about 17.32 words. While Penny participates in 5,566 conversation turns, each containing about 12.05 words. The average utterance audio duration per character reaches 5.93h, average utterance number per character reaches 9,717.8, greatly exceeding all the existing multimodal datasets. This relatively large volume of character data makes it possible to model personality through multimodal inputs. To be specific, as we offer sufficient utterance data and conversational data, it is possible to model the character's speaking manner and interaction habits. With hours of audio tracks, one can capture the character's voice and tone to build a personalized speech synthesis model. One can also rebuild the character's visual image or 3D model from the video clips, and analyze the character's facial expression tendency during the conversation.

4. Our DPCC Baseline

In this section, we provide a baseline for DPCC, and the whole pipeline is presented in Figure 3.

Multimodal Features Encoding. We train a multimodal feature representation model to capture each modality's distinctive stylistic information. Following MISA[13], one-dimensional semantic features are previously extracted from each utterance or audio/video clip, then projected into a modality-specific space utilizing feature projection model E_p. E_p is pretrained on a multimodal emotion recognition task[3, 48] for better understanding in multimodal correlation. The textual/acoustic/visual features are encoded by E_p to form multimodal conversation context input $u_{t,a,v}$.

ICCV
#3137

ICCV
#3137

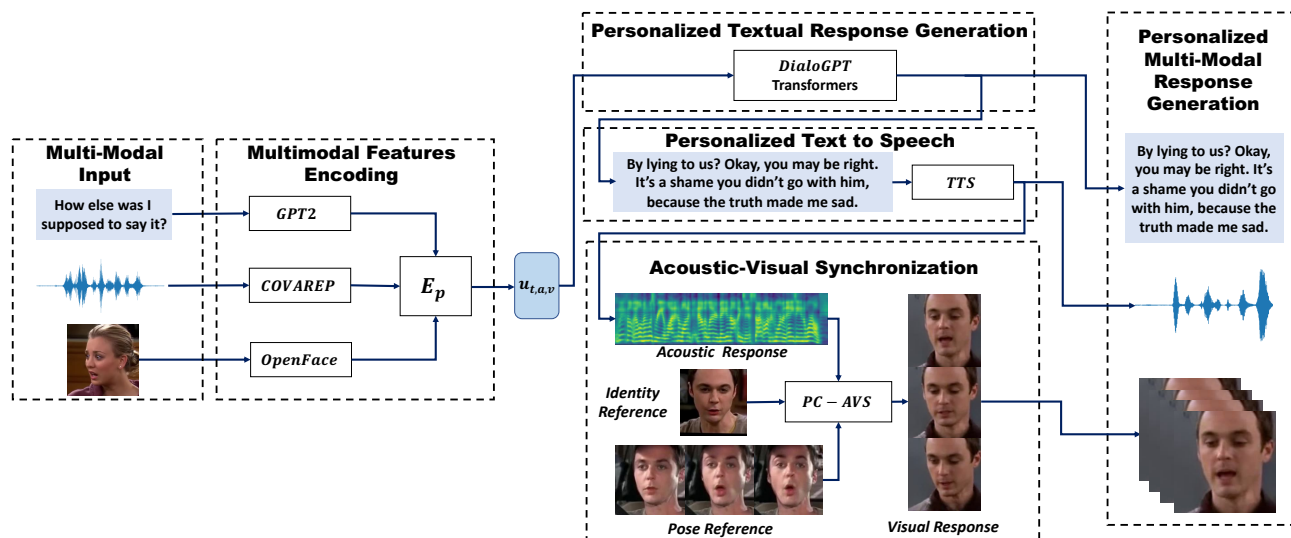ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Figure 3. Overview of our baseline solution for DPCC. Given multimodal input, we first extract multimodal features, then the personalized textual, speech and video responses are generated by a personalized textual response generation model, a personalized text-to-speech model, and an acoustic-visual synchronization model, sequentially and respectively.

Personalized Textual Response Generation. We fine-tune the pretrained model conversational response generation DialoGPT[51] on a single character's multimodal conversation context input $u_{t,a,v}$ to model his personalized interaction pattern and speaking style and finally generate the specific character's possible response to the multimodal input context.

Personalized Text to Speech. We use the TTS model proposed by Jia et al.[16] to utter the generated textual response in the target character's voice and tone. For each character, the TTS model is finetuned on wave files and utterances collected from that specific character to capture unique speech patterns.

Acoustic-Visual Synchronization. Utilizing the acoustic-visual synchronization model, we synthesize our generated textual response and corresponding audio to the final multimodal response. We adopt the pretrained Pose-Controllable Talking Face Generation Model(PC-AVS) proposed by Zhou et al[54]. The identity reference and pose reference are randomly selected from the target character's video clips and lip movements are synchronized with the wave file generated by the personalized speech synthesis model. In this way, all generated modalities are aligned and synthesized to a final multimodal personalized response.

5. Experiments

5.1. Experimental setup

Baseline. We use the multimodal data collected in our dataset to train the baseline DeepCharacter, and separate models are trained for different characters, respectively. To pretrain multimodal feature representation models, we utilize multimodal features and emotion labels from CMU-MOSEI dataset[3]. We set the batch size to 8 and hidden feature size to 64, and use the original default settings. The encoded 64-dimension audio/vision features are concatenated with 128-dimension word embeddings transformed by pretrained GPT2 tokenizer. Then we use this multimodal input to finetune DialoGPT-small conversation response generation model. See Supplement Section 3 for detail setting. As for the personalized speech synthesis model, the audios are transformed to 22050Hz sample rate first and aligned with utterances using Montreal Forced Aligner(MFA)[5]. Videos should be transformed to 25fps and wavfiles to 16kHz for Pose Controllable Talking Face Generation, and all other settings by default. All models are implemented by PyTorch[23] and performed on a RTX-3090 GPU.

Evaluation. Since our proposed task and baseline method heads towards creating personalized characters, the degree of characterization is a key benchmark to evaluate the model performance. We evaluate both the characterization and quality. In spite of objective experiments, we also conduct subjective experiment to evaluate the user experience. The objective evaluation is on all the 5 DeepCharacters: DeepSheldon, DeepPenny, DeepLeonard, DeepHoward and DeepRaj. For the subjective evaluation, we conduct a survey on 50 people and choose 30 judges according to their knowledge of the characters, see Supplement Section 2.1 for the standard details. The survey shows that three characters Sheldon, Leonard and Penny, are more impressed and related conclusion should be more

---

[5]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ICCV
#3137

	DeepSheldon	DeepPenny	DeepLeonard
C1: Dammit man, we're dying here.	No we aren't. This is a swirling vortex of entropy.	Yeah, I know.	What are you doing?
C2: You know, if you guys are interested, there's a technique where I can take a lock of your hair, refine it into carbon dust and use the hydraulic press at work to turn it into a tiny little diamond with your DNA in it.	I'm sorry, but that'd be impossible. The reaction time would be too much.	Oh my God, that is so stupid.	That sounds fun.

Table 4. The response of different characters to same textual input.

reliable, thus we conduct subjective evaluation on the corresponding 3 DeepCharacters: DeepSheldon, DeepPenny and DeepLeonard. Since the textual response lays the foundation of multimodal response, and psychologists think that language can express inner thoughts and feelings[1], so it's possible to distinguish a specific character according to his/her utterances[42]. Therefore, we conduct experiments on both textual and multimodel ones.

5.2. Characterization Evaluation

Characterization of textual responses. As a vivid illustration, we present 2 samples of different characters response to same textual inputs in Table 4. Character's personality can be somehow reflected in these different responses: Sheldon is a high IQ scientist who uses scientific terms like 'vortex of entropy' or 'reaction time' from time to time; Penny, on the other hand, is a woman with rich affection, preferring interjection like 'Yeah','Oh god' or 'Sweetie'; Leonard, although also being somewhat bookish, is more like a normal and helpful person who questions strange utterances and tend to give favorable replies. And as proved by characterization clustering, character classification and human evaluation, our DeepCharacter models can generate well-characterized textual responses.

Following a previous work [38] on character classification, we use SAGE[9] model to derive weights for words uttered by the characters. SAGE enforces a sparse prior on its parameters, and may be sensitive to infrequent terms in the text. To alleviate this issue, given one character, we experimentally group 100 randomly selected responses from this character, forming a virtual document[39], and reweight the words inside the document utilizing SAGE. Then we use a pretrained basic BERT model to encode the reweighted virtual documents into 128-dimensional embeddings as the final representation of one sample of the given character's response. We perform t-SNE[37] on these sampled representations from all the 5 charaters and 5 DeepCharacters. Note that during the above process, BERT are fixed. See details in Supplement Section 3.3 for implementation details.

As presented in Figure 4, the responses from 5 characters can be well separately clustered, showing the character's diversity in terms of speaking style and word preference. Intriguingly, the clusters of responses DeepCharacters aligns quite well with the ones of the corresponding characters, respectively. This verifies that the created DeepCharacters can generate highly personalized textual responses,
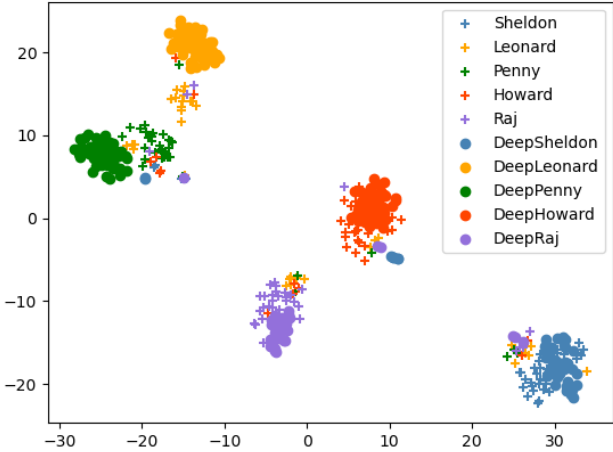

Figure 4. t-SNE visualization of characters' original responses and DeepCharacters' generated ones. Different colour represents different characters. Cross and circle mark represents original and generated text, respectively.

well capturing the speaking style and word preference.

To provide quantative results, we also build a character classifier based on BERT to automatically evaluate the characterization of text responses generated by our DeepCharacters. The character classifier is trained on 18,000 documents from original corpus. See details in Supplement Section 3.3 for implementation details. This character classifier achieves 98.3% validation accuracy on origin corpus (900 documents) while getting 86.4% testing accuracy on generated responses (2,000 documents), showing that our DeepCharacters can well capture the characteristics. More discussions on utterance grouping number and more visualization results are provided in Supplement Section 4.2.

For human evaluation, we randomly select 50 response samples from DeepSheldon, DeepPenny and DeepLeonard respectively, and each is paired with general response generated by DialoGPT as a non-personalized baseline, and response generated by other character model as an untargeted character comparison. Each question is presented to at least 3 judges.

The judges are asked to rank each response pair for how well the response matches the given material in speaking style, decision-making style and emotion tendency, using a 3-point Likert-like scale following [51]. As shown in Table 5, all three targeted DeepCharacters gain more preferences over non-personalized DialoGPT and untargeted DeepCharacters, which shows our DeepCharacter's ability of capturing a specific character's personality and speaking style.

Characterization of multimodal responses. As characterization in multi-modality is comparatively subjective, we rely on user study to evaluate our DeepCharacter's capacity in generalizing personalized multimodal responses.

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ICCV
#3137

Model	Sheldon	Penny	Leonard
DialoGPT	0.420	0.058	0.439
Neutral	0.077	0.321	0.058
DeepCharacter	**0.503***	**0.622***	**0.503***
Other DeepCharacters	0.367	0.342	0.465
Neutral	0.042	0.065	0.026
DeepCharacter	**0.590***	**0.594***	**0.510***

Table 5. Results of Human Evaluation for characterization in textual responses, showing preferences for our model: DeepCharacter vis-a-vis DialoGPT or other DeepCharacter model. Numbers in bold indicate the preferred models. Here, *other DeepCharacters* refers to DeepCharacter trained on characters other than the target one, and *Neutral* refers to neutral option without preference for either model. Statistically significant responses are indicated: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Multimodal Responses	Model	Sheldon	Penny	Leonard
Multimodal Quality	Baseline	0.29	0.33	0.29
	DeepCharacter	**0.71***	**0.67***	**0.71***
Characterization	Baseline	0.10	0.32	0.16
	DeepCharacter	**0.90***	**0.68***	**0.84***

Table 6. Results of Human Evaluation for multimodal responses, showing preferences for our model: DeepCharacter vis-a-vis neutral baseline. Numbers in bold indicate the preferred models.Statistically significant responses are indicated: *** $p \leq 0.001$.

Model	Dist-1		Dist-2		Avg.len		Entropy-4		Perplexity		SacreBleu
	O-M	R-D	O-M	R-D	O-M	R-D	O-M	R-D	O-M	R-D	
DeepSheldon	0.413	0.465	0.841	0.902	14.01	14.90	7.68	7.76	172.687	174.365	0.510
DeepPenny	0.341	0.401	0.798	0.882	10.18	12.37	7.23	7.53	131.339	144.014	1.850
DeepLeonard	0.383	0.422	0.813	0.892	9.56	10.32	7.16	7.28	142.605	167.472	1.183
DeepHoward	0.422	0.442	0.855	0.897	13.42	14.08	7.512	7.597	176.661	177.023	0.503
DeepRaj	0.423	0.448	0.874	0.917	13.66	14.21	7.511	7.616	241.974	215.250	0.345
DialoGPT	0.217		0.414		7.955		6.33		34.561		0.160
DialoGPT (finetuned)	0.384		0.794		11.475		7.25		163.76		0.254

Table 7. Automatic evaluation scores including O-M (Our Method) and R-D (Reference Data).

We use DialoGPT finetuned on whole DPCD corpus instead of any specific character, and un-finetuned TTS model as the neutral baseline, while the inference of acoustic-visual synchronization model remains the same as DeepCharacter. The neutral baseline has learned the TV show corpus's special text distribution and can generate high quality speeches.

We randomly select 10 video clips generated by DeepSheldon, DeepLeonard and DeepPenny, respectively. Like PC-AVS[54], the total 30 video clips, paired with baseline results, are handed to all 30 judges. The judges are asked to rank each multimodal response pair base on how well the response matches the target character in terms of speaking style, tone and identity, considering all three modalities comprehensively. As shown in Table 6 , all three targeted DeepCharacters gain more preferences over the neutral baseline, which shows our DeepCharacter's ability of generating characterized multimodal responses.

The above results demonstrate the collected DPCD can support our simple baseline to generate deep personalized characters.

5.3. Quality evaluation

We perform automatic and human evaluations on the quality of responses generated by our DeepCharacter model. Here, high quality responses refer to reasonable texts, realistic speeches and videos, and of high naturalness and consistency across all three modalities.

Quality of textual responses. Following DialoGPT[51], we perform automatic evaluations using several popular standard evaluation metrics, including ScareBLEU[28] and Perplexity[15]. SacreBLEU provides hassle-free computation of BLEU scores, ranging from 0 to 100, and a higher SacreBLEU means a better match between generated results and references. Perplexity measures how likely the model is to generate the input text sequence and can be used to evaluate how well the model has learned the distribution of the training text. We also use Entropy[50] and Dist-n[17] to evaluate lexical diversity.

We calculate automatic evaluation scores for DeepSheldon, DeepPenny, DeepLeonard, DeepHoward and DeepRaj separately. In comparison, we evaluate DialoGPT's results with the same contexts, as a non-personalized general response. We also finetune DialoGPT on the mixed corpus of these characters, to bridge the performance gap caused by corpus difference and serve as an un-targeted response. We additionally evaluate the lexical diversity and perplexity of the characters' original corpus, to learn text distribution.

Table 7 summarizes the automatic evaluation results, showing DeepCharacter's stronger adaptability to complicated daily problems. All DeepCharacters achieve a higher score in lexical diversity than original DialoGPT, and are close to respective original corpus. Both generated responses and original corpus have higher perplexity than DialoGPT's results, illustrating the complexity and distribution diversity of our DPCD, which can also be verified by the results of finetuned DialoGPT. This also demonstrates DeepCharacter's ability to fit text distribution of the original corpus. Due to high complexity and ambiguity of generation task, all five DeepCharacters receive rather low SacreBLEU score, but still higher than original and finetuned DialoGPT, showing a stronger reasoning ability to understand and reply diverse and colloquial dialogues.

We also conduct human study to evaluate our model's ability in generating high quality characterized textual responses. The testing data remain same with characterization evaluation. We first evaluate the generated text quality from the views of fluency, context relevant and personalized/emotional degree. Judges rate between 1 to 3 for each aspect, where 1 means "very bad, not at all fluent/not at all relevant/very general", 2 means "not so bad, some minor mistakes", and 3 means "very fluent/relevant/personalized or emotional".

As shown in Table 8, all DeepCharacters reach high scores on all three evaluation criteria. According to signif-

ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Elements	Model	Sheldon	Penny	Leonard	Avg.of Model
Fluency	DialoGPT	-	-	-	2.646
	DeepCharacter	2.720	2.602	2.630	**2.651**[+]
Context Relevance	**DialoGPT**	-	-	-	**2.169**[+]
	DeepCharacter	2.001	2.240	2.216	2.152
Emotional Degree	DialoGPT	-	-	-	2.066
	DeepCharacter	2.048	2.161	2.090	**2.100**[+]

Table 8. Results of Human Evaluation for text quality, showing preferences for our model: DeepCharacter vis-a-vis DialoGPT. The average text quality score of our three DeepCharacters is calculated and compared with other models in the last column, and numbers in bold indicate the preferred models. Statistically significant responses are indicated: [+] $p \geq 0.5$

icance test, the differences between DeepCharacter and DialoGPT in these three criteria are less significant, indicating that our DeepCharacter can generate responses that are as fluent, reasonable, and emotional as DialoGPT. While DialoGPT is well-known for generating high quality responses given basic and general contexts, our DeepCharacter can achieve similar quality in complicated personalized scenarios.

The above results demonstrate that our DeepCharacters achieve comparatively high score in lexical diversity, reasonalibty, fluency and emotionality of generated texts.

Quality of generated multimodal responses. We conduct user study to evaluate the multimodal quality of the generated videos. The data and experiment settings remain the same with the video characterization user study, and the judges are required to rank each multimodal response pair for the responses' performance in language fluency, naturalness and realness.

As shown in Table 6, all three DeepCharacters gain more preferences over comparative baseline, which confirms our DeepCharacter's ability of generating high quality multimodal responses. Although videos generated by both models shares similar speech clarity and image sharpness, the vivid tone and reasonable textual responses generated by our DeepCharacter increase naturalness and realness, resulting in judges' preference when considering multimodal quality. Please see the Supplementary Materials for the multimodal survey contents.

To study how the multimodal input affects the response, we take the same textual stimuli mixed with different video/audio context. As shown in Figure 5, instead of generating the same response to the shared text, our model gives different results not only in response to the textual context, but also related to the emotion expressed in the video/audio context. Specifically, our DeepCharacter tends to generate a more tender or more positive response when given video/audio inputs expressing sad moods. On the contrary, when the speaker in the video/audio input is very angry and unfriendly, the model will be more likely to generate unkind words, sometimes seeming to fight back, like

Figure 5. Examples of DeepSheldon response with multimodal inputs. For each row, given the inputs of the same textual context and but different video/audio context with different emotions (labelled with different colors).

"*How am I supposed to know that?*", or "*No, I didn't get it, I thought it was funny. Apparently neither did you.*" Extra modalities can offer more effective information, such as the speaker's facial expressions, voice or tone, which can better express the speaker's emotion and will certainly affect natural response generation. Given multimodal inputs, the model can obtain more comprehensive information from the conversation context, thus understanding the character's interaction pattern and personality better, generating diverse and more personalized responses.

6. Conclusion

Towards a fantastic goal of creating personalized characters with whom one can interact in a multimodal channel, we explore creating personalized characters from collected multimodal data, in a data-driven way. In this paper, we formulate a novel task (DPCC), given multimodal-in stimuli, and predict multimodal-out response. To support this task, we collect a multimodal conversation dataset (DPCD) surrounding several characters from TV shows. This dataset provides about 10 times more multimodal conversations of high quality, well aligned across multimodality. We further provide a baseline solution for learning multimodal response given multimodal input. The created DeepCharacters is capable of generating relatively high-quality response, and consistent personality to the original character. In the future, we would like to collect richer characters' data such as pose, gestures etc, and more rounds of conversations to further enrich the personalized feature. We suppose this work can shed light on an interesting research direction towards creating deep personalized characters to support multimodal chatting scenarios, and inspire further works on, such as, improving the visual and acoustic quality, proposing metrics for evaluating personality similarity and fidelity, extending the 2D video to 3D for more immersive interaction.

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ICCV
#3137

References

[1] Hussain Ahmad, Muhammad Zubair Asghar, Alam Sher Khan, and Anam Habib. A systematic literature review of personality trait classification from textual content. *Open Computer Science*, 10(1):175–193, 2020. 6

[2] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning (ICML)*, pages 195–204. PMLR, 2017. 1

[3] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3, 4, 5

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Jeannette N. Chang, Samuel Kim, Sungbok Lee, and Shrikanth S. Narayanan. IEMO-CAP: interactive emotional dyadic motion capture database. In *Language Resources and Evaluation*, pages 335–359, Dec. 2008. 2, 3, 4

[5] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. 2, 3, 4

[6] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Adaspeech: Adaptive text to speech for custom voice, 2021. 3

[7] Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*, 2022. 2, 4

[8] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 87–103, Cham, 2017. Springer International Publishing. 3, 4

[9] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1041–1048, Madison, WI, USA, 2011. Omnipress. 6

[10] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4441–4453, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 2, 3, 4

[11] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*, 13(3):1555–1566, 2022. 2, 3, 4

[12] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. 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)*, pages 2046–2056, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 3, 4

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. 4

[14] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 3, 4

[15] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977. 7

[16] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4485–4495, Red Hook, NY, USA, 2018. Curran Associates Inc. 5

[17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. 7

[18] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 2

[19] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1, 2

[20] Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. MSCTD: A Multimodal Sentiment Chat Translation Dataset. *arXiv e-prints*, page arXiv:2202.13645, Feb. 2022. 3, 4

[21] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dia-

9

ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

logue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2775–2779, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[22] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 3, 4

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019. 5

[24] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021. 1

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[26] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning (ICML)*, pages 8599–8608. PMLR, 2021. 1

[27] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. 2, 3, 4

[28] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 7

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 1

[31] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2020. 3

[32] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[36] 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. 1

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 6

[38] Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. Are fictional voices distinguishable? classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34, Minneapolis, USA, June 2019. Association for Computational Linguistics. 6

[39] Krishnapriya Vishnubhotla and Saif M. Mohammad. Tweet emotion dynamics: Emotion word usage in tweets from us and canada. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, 2022. 6

[40] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020. 1, 2

[41] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 700–717, Cham, 2020. Springer International Publishing. 3, 4

[42] Zhe Wang, Chun-Hua Wu, Qing-Biao Li, Bo Yan, and Kang-Feng Zheng. Encoding text information with graph convolutional networks for personality recognition. *Applied Sciences*, 10(12), 2020. 6

[43] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia (ACMMM)*, pages 1478–1486, 2021. 1, 2

[44] Yuwei Wu, Xuezhe Ma, and Diyi Yang. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(ACL-HLT)*, pages 1956–1970, Online, June 2021. Association for Computational Linguistics. 2

ICCV
#3137

ICCV
#3137

ICCV 2023 Submission #3137. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

[45] Yujie Xing and Raquel Fernández. Automatic evaluation of neural personality-based chatbots. In *Proceedings of the 11th International Conference on Natural Language Generation (NLG)*, pages 189–194, Tilburg University, The Netherlands, Nov. 2018. Association for Computational Linguistics. 2

[46] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). https://doi.org/10.7488/ds/2645., 2019. 3, 4

[47] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3718–3727, Online, July 2020. Association for Computational Linguistics. 3, 4

[48] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 4

[49] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 3, 4

[50] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS 2018*, December 2018. 7

[51] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics. 3, 5, 6, 7

[52] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672, 2019. 2

[53] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online, Nov. 2020. Association for Computational Linguistics. 2

[54] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4174–4184, 2021. 5, 7