EmoteGPT: 3D Human Facial Expressions from Natural Language Descriptions

Haoran Wang[1], Mohit Mendiratta[1],
Christian Theobalt[1], and Adam Kortylewski[2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] CISPA Helmholtz Center for Information Security

Project Page: `https://genintel.github.io/EmoteGPT/`

Abstract. Precise control of 3D facial expressions from text is crucial for virtual avatars, animation, and human–computer interaction, yet existing text-to-3D methods jointly generate identity, expression, and texture, making fine-grained expression control difficult. We instead formulate text-driven expression synthesis as a regression problem in the disentangled parameter space of a 3D Morphable Model (3DMM). This setting, however, requires paired data linking detailed language to precise expression parameters, which is missing from existing resources. To fill this gap, we introduce Txt2Emote, a benchmark of diverse 3D facial expressions with fine-grained textual annotations obtained from GPT-4o and a high-fidelity face tracker, providing both explicit descriptions detailing facial features and implicit descriptions referencing the situational context behind the expression. Leveraging this dataset, we present EmoteGPT, a text-to-3D expression framework based on a Multi-Modal Large Language Model (MLLM) with a dedicated `<Expr>` token to semantically ground expression representations, which are then decoded into 3DMM parameters. We further improves EmoteGPT by augmenting training with large-scale image-to-3DMM data, surpasses state-of-the-art text-to-3D face synthesis methods on emotion recognition metrics and in perceived expressiveness. Integrated into avatar pipelines, our method enables photorealistic and stylized 3D avatars, as well as expressive 3D-consistent 2D face synthesis from textual input.

Keywords: Visual Language Model · Face Synthesis

1 Introduction

Controllable 3D facial expression has broad applications in virtual avatars, animation, and human–computer interaction. Traditional control methods such as blendshape editing are labor-intensive and unintuitive for non-experts. In contrast, natural language offers a flexible and universal interface for expression control. Users can describe expressions in diverse ways, using **explicit visual**

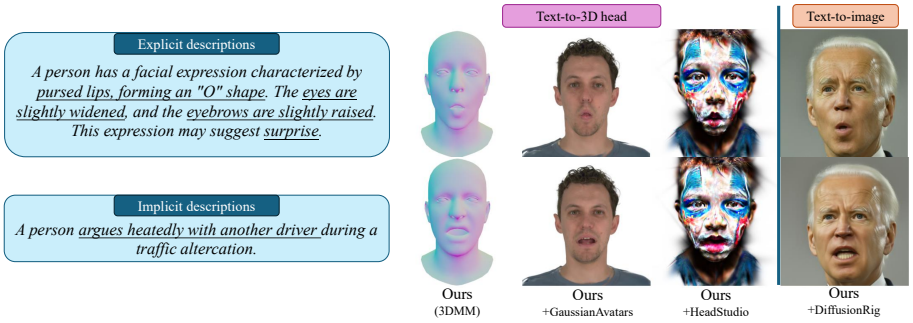

Fig. 1: EmoteGPT generates 3D facial expressions from text, supporting two types of descriptions: (1) *Explicit* ones, which detail physical facial features and overall emotional impressions; (2) *Implicit* ones, which reference events or situations that evoke facial expressions. The expressions are represented as 3DMM parameters and can be seamlessly integrated into existing frameworks [28, 43, 48] to enable expressive 3D avatars and personalized 2D face synthesis.

cues (e.g., "He is smiling") or **implicit contextual statements** (e.g., "He looks like he just won the lottery"). Explicit descriptions specify observable facial features, while implicit ones encode the situational or emotional context that evokes those features. Handling both forms is crucial for natural interaction, as people rarely describe emotions by enumerating muscle movements.

Recent works [3, 23] have explored generating expressive 3D heads from text, predominantly using diffusion-based text-to-3D pipelines. However, these methods remain limited for 3D expression control: they typically entangle identity, pose, and expression. They also struggle to capture nuanced, diverse language descriptions, and require expensive iterative sampling that limits efficiency. As a result, existing text-to-3D head generation methods fall short in scenarios demanding accurate, independently controllable, and efficient expression synthesis.

We instead address these challenges by formulating the problem as a *language-to-expression regression* task, where expressions are represented explicitly in the compact and disentangled parameter space of a 3D Morphable Model (3DMM) [2, 6]. This 3DMM-based formulation provides a structured and interpretable representation that decouples expression from identity, enhances controllability and expressive fidelity, and enables efficient inference.

A key obstacle is the lack of data linking natural language to precise 3D expression parameters. To address this, we construct Txt2Emote, a dataset of 30k 3DMM expressions paired with diverse textual descriptions. Each expression is annotated with two complementary text types: (1) **explicit** annotations that detail observable facial features (e.g., mouth shape, eye tension), and (2) **implicit** annotations reflecting emotional or situational context. Figure 1 illustrates representative samples. We also reserve 2.5k expressions as a held-out benchmark for evaluating models on both explicit and implicit language.

Building on Txt2Emote, we propose EmoteGPT, a framework for generating 3D facial expressions from natural language. EmoteGPT uses a Multimodal Large Language Model (MLLM) to directly regress 3DMM expression parameters. We introduce a dedicated `<Expr>` token whose representation can be decoded by a lightweight expression head into 3DMM parameters, yielding deterministic, real-time predictions without diffusion-style iterative sampling. Operating in the 3DMM expression space allows EmoteGPT to focus exclusively on expression, enabling precise and controllable manipulation while remaining compatible with existing avatar systems.

EmoteGPT further supports multimodal supervision, leveraging large-scale image-to-expression data and instruction-following text data to strengthen the alignment between language and facial expression. The resulting model integrates seamlessly with existing 3DMM-based avatar pipelines [28, 38, 48], enabling a wide range of applications including photorealistic avatar synthesis, stylized avatar generation, and personalized facial modeling (Figure 1). Extensive experiments demonstrate that EmoteGPT significantly outperforms prior CLIP- and diffusion-based baselines in both emotion recognition accuracy and expressive fidelity.

Our contributions are summarized as follows:

- We propose EmoteGPT, a new method that leverages a Multimodal Large Language Model (MLLM) for generating 3D facial expressions from natural language, enabling precise and expressive facial expression synthesis.
- We introduce Txt2Emote, a new dataset of 30k 3D facial expressions paired with fine-grained textual annotations, including *explicit* visual cues and *implicit* contextual texts.
- We demonstrate that EmoteGPT benefits from multimodal face data to strengthen the link between language and facial expression and that it substantially outperforms state-of-the-art CLIP-based diffusion models and MLLM baselines on both emotion recognition accuracy and expressiveness.

2 Related works

Face dataset with text annotations. Several multimodal face datasets with text annotations exist, but most predominantly focus on describing static visual attributes or general appearance, offering only limited coverage of facial expressions. CelebA-Dialog [13] provides captions of five attributes for face images in the CelebA dataset where only the smile attribute is related to human facial expressions. MMCelebA-HQ [42] uses an automatic template-based approach to generate captions from attribute annotations in CelebA-HQ, limiting texts' diversity and expressiveness. CelebAText [33] provides human-labeled captions for 15k images in CelebA-HQ dataset, yet the annotations primarily emphasize high-level visual attributes or overall appearance rather than nuanced expression details. While some datasets include expression-related text annotations, such descriptions tend to be coarse, capturing only basic states, e.g. whether the subject is smiling or has an open mouth, without detailing the underlying facial

muscle movements or compound emotions. As a result, current multimodal face datasets with textual descriptions remain insufficient for studying or modeling fine-grained facial expressions.

Such limited annotations hinder the ability of models to learn fine-grained mappings between language and facial expression. Furthermore, existing datasets often lack contextual or situational cues that might evoke expressions. This absence of contextual grounding makes it challenging for generative models to synthesize realistic and expressive faces from text, particularly when the input does not explicitly describe facial expressions.

Text-to-3D face generation. Early works on text-to-3D face synthesis, such as Describe3D [39], CLIP-Face [1] and CLIP-Head [22], builds on CLIP to achieve text-guided 3DMM face synthesis by aligning the visual features of rendered faces with the semantic meaning of text prompts. Inspired by DreamFusion [26], which demonstrated the potential of Score Distillation Sampling (SDS) for 3D object generation, recent methods [12, 41, 48] have started leveraging large pre-trained text-to-image diffusion models for 3D face synthesis from text. However, SDS-based 3D generation often suffers from saturation and noisy artifacts, prompting efforts to address these limitations. For example, HumanNorm [12] introduces a multistep SDS pipeline for progressive 3D human mesh generation, while Portrait3D [41] incorporates a 3D-aware GAN to reduce errors, and HeadStudio [48] uses the FLAME model to improve geometric initialization. Despite these improvements, these methods primarily focus on shape identity and overlook expressive facial details. Some audio-based head generation works [5, 21, 34, 45] incorporate text as additional guidance for expressive synthesis with 3DMMs, but they typically depend on emotion labels or handle only limited, simple descriptions. In contrast, EmoteGPT explicitly bridge diverse complicated text description to various expressive 3D face parameters.

Multi-modal Large Language Models. Multi-Modal Large Language Models extend the capabilities of language models beyond text to include a broader spectrum of modalities, such as images, videos, and audio. Recent research in this area enhances their ability to process and integrate multiple modalities, making these models more versatile. In the realm of image-text understanding, approaches like LLaVA [18] and MiniGPT-4 [49] incorporate vision encoders to interpret images and align their features with language embeddings using projection layers. Furthermore, models such as PandaGPT [32], ImageBind [9], and NeXT-GPT [40] show versatility in handling diverse modalities, aligning embedded text, images, audio and video with language as input and output. Approaches such as LISA [15] connect LLaVA with a decoder to generate text and segmentation masks, while ChatPose [8] specializes in human pose information. However, there has been no study on integrating MLLMs with 3D human facial expressions, and our work fills this gap.

Image based expressive 3D face reconstruction. Expressive 3D face reconstruction has been widely explored in the context of image-based approaches. Existing works [4, 31] mainly take single image as input and output 3D facial expression in FLAME head model space [16]. EMOCA [4] incorporates emo-

tion consistency and lip articulation constraints to refine expression accuracy, while SMIRK [31] extends MICA [50] by leveraging an image-to-image translation network for improved expression synthesis. These image-based 3DMM estimation works have been widely integrated in 2D, 3D or 4D avatar synthesis pipelines [27, 35, 36, 43]. While these methods focus on reconstructing expressive 3D faces from images, generating expressive 3D faces directly from text descriptions remains less explored.

3 Text-to-3D expression Dataset

Aligned text descriptions for diverse facial expressions are necessary to establish a connection between language and 3D facial expressions. However, few public datasets contain paired text with 3D faces. As mentioned in Section 2, existing face datasets with texts [13,33,42,47] provide limited text captions for facial expressions, which hinders the research on text-to-3D facial expression generation. To fill this gap, we introduce Txt2Emote, a new dataset of expressive faces with individual descriptive captions and identity-isolated 3D meshes.

Table 1: Comparison of our dataset with existing facial-expression datasets containing descriptive text. *Partial* indicates that descriptions mention only coarse cues (e.g., smiling, open mouth) without covering other regions such as the eyes or nose, or conveying the overall emotional context. Our dataset offers fine-grained, localized descriptions and additional *implicit* texts describing scenarios likely to elicit the expressions.

Dataset	Samples	Indiv. Expr.	Explicit	Implicit
MMCelebA-HQ	30k	✓	partial	✗
CelebAText-HQ	15k	✗	partial	✗
FFHQ-Text	760	✗	partial	✗
CelebA-Dialog	202k	✓	partial	✗
Txt2Emote (Ours)	30k	✓	✓	✓

3.1 Dataset Construction

We construct our dataset from AffectNet dataset [24] which contains diverse expressive face images spanning eight emotions. Importantly, we follow the intuition that humans describe facial expressions in two main ways: (1) *direct* descriptions for facial attributes, such as movements of eyes, nose or mouth and the overall emotions (e.g. “The face has a smile. This face looks happy"), and (2) *indirect* descriptions via activities or scenarios that could evoke the expression (e.g “The face looks as if he/she just failed an exam"). In this work, we refer to the first type as *explicit* descriptions and the second as *implicit* descriptions. For

each expressive facial instance in our dataset, we provide a 3D face mesh along with both forms of text:

Explicit descriptions: These provide comprehensive details of facial features along with a high-level summary of the associated emotion. Unlike existing datasets, which offer only coarse or partial annotations, our descriptions capture the comprehensive complexity of expression by specifying physical attributes in different facial regions and emotional nuance.

Implicit descriptions: In contrast to existing face datasets, which focus solely on facial features or emotion labels, implicit descriptions depict everyday scenarios or actions that naturally evoke the corresponding facial expression. This offers a more natural, human-centric way of describing expressions and aligns with how people intuitively understand and communicate emotions.

To build the dataset, we first subsample AffectNet according to its emotion annotations to obtain a balanced set with a similar number of images per emotion category. For each selected image, we reconstruct a 3D facial mesh using a robust expressive face estimator [4] and generate both explicit and implicit descriptions with GPT-4o [25]. We then filter the reconstructed meshes and generated text annotations to reduce noise. The resulting dataset contains 30k expressive facial images, each paired with a 3D mesh and two text annotations. Among them, 27k samples are drawn from the AffectNet training set and 3k from the AffectNet test set. Following the AffectNet protocol, we manually inspect the 3k test samples and retain 2.5k high-quality samples as the evaluation benchmark. Table 1 compares our dataset with existing face datasets containing text annotations. Further details on data generation and prompting are provided in Supplementary Sec. A.

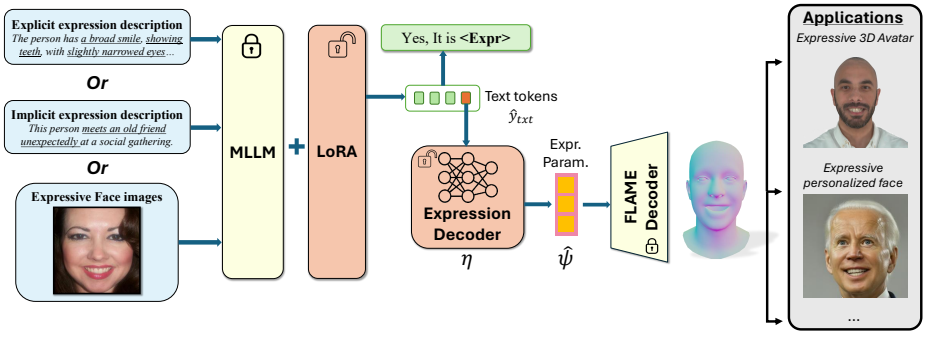

Fig. 2: Overview of EmoteGPT. Our framework generates 3D facial expressions from diverse inputs by combining a Multi-Modal Large Language Model (MLLM) with an expression decoder head ($\boldsymbol{\eta}$). EmoteGPT is trained on diverse multi-modal supervision, including explicit textual descriptions (e.g., "a broad smile with narrowed eyes"), implicit contextual prompts (e.g., "meeting an old friend unexpectedly"), and images. During training, η is optimized, and the LLM within the MLLM is fine-tuned using LoRA while other components remain frozen. The generated expressions can be seamlessly integrated with FLAME-based face avatars for diverse applications.

4 Method

In this section, we describe the EmoteGPT framework for generating 3D facial expressions from natural language descriptions. Our approach combines a Multi-Modal Large Language Model (MLLM) with a lightweight expression decoder to establish a robust mapping from text to 3D facial expression parameters. We leverage the FLAME model [16] as the face representation, enabling precise control over facial expressions in a structured parameter space. In Section 4.1, we introduce the facial expression parameterization based on FLAME. Section 4.2 details the overall EmoteGPT pipeline, including the use of the `<Expr>` token and its integration within the MLLM. Finally, Section 4.3 outlines the multi-modal training strategy, including supervision from both text and image data, and the overall optimization objective.

Table 2: Quantitative results for text-to-3D facial expression generation with different descriptions. The best performance is highlighted with **bold** font. EmoteGPT outperforms other baselines across explicit and implicit benchmarks. EmoteGPT also shows the best generalization ability even when trained without implicit descriptions, maintaining strong performance on explicit and unseen inputs.

Method	Training		Explicit Expression Descriptions						Implicit Expression Descriptions					
	Exp	Imp	V-CCC ↑	A-CCC ↑	E-ACC↑	$L1_{Head}$ ↓	$L1_{Face}$ ↓	$L1_{Lip}$ ↓	V-CCC ↑	A-CCC ↑	E-ACC↑	$L1_{Head}$ ↓	$L1_{Face}$ ↓	$L1_{Lip}$ ↓
FLAME			-	-	-	0.83	1.41	3.53	-	-	-	0.83	1.41	3.57
Describe3D			0.48	0.34	0.34	-	-	0.19	0.08	0.19	-	-		
CLIP	✓		0.61	**0.57**	0.51	0.66	1.06	2.46	0.36	0.28	0.32	0.84	1.45	3.70
Vicuna-1.5-7b	✓		0.56	0.53	0.51	0.66	1.08	2.48	0.47	0.34	0.38	0.80	1.35	3.37
EmoteGPT	✓		**0.61**	0.50	0.51	**0.61**	**1.00**	**2.36**	**0.49**	**0.44**	**0.41**	**0.74**	**1.30**	**3.27**
CLIP	✓	✓	0.64	0.56	0.52	0.65	1.07	2.50	0.49	0.41	0.41	0.73	1.24	3.04
Vicuna-1.5-7b	✓	✓	0.60	0.55	0.52	0.65	1.08	2.50	0.49	0.39	0.41	0.71	1.21	2.94
EmoteGPT	✓	✓	**0.68**	**0.62**	**0.59**	**0.60**	**0.98**	**2.29**	**0.57**	**0.48**	**0.44**	**0.66**	**1.09**	**2.69**

4.1 Facial Expression Representation

We adopt the 3D Morphable Model (3DMM) framework [2] to represent 3D head geometry with disentangled identity and expression parameters, enabling structured and controllable manipulation.

Specifically, we leverage FLAME [16], a statistical head model parameterized by identity shape β, facial expression ψ, and pose parameters θ for rotations around four joints (neck, jaw, and eyeballs), and the global rotation. FLAME effectively disentangles facial expressions from identity and head pose. In EmoteGPT, we focus on predicting expression parameters (ψ) and jaw rotation (θ_{jaw}), while keeping identity (β) and head pose (θ_{head}) fixed at zero by default. This ensures that generated expressions are consistent, disentangled, and placed on an average-sized, neutrally posed head model, i.e., identity-isolated.

4.2 Pipeline Overview of EmoteGPT

To establish a robust and generalizable mapping from language to facial expressions, our approach requires a model with strong text understanding capabilities. A core design choice in our model is to build on a powerful multi-modal large language model (MLLM) backbone. Building EmoteGPT on a MLLM enables the model to learn from diverse facial expression data including text-to-3D-face and image-to-3D-face data. As a result, our system can efficiently utilize multi-modal inputs (as demonstrated in our experimental results in Section 5.2).

Representing facial expression in language space. To build a connection between MLLM and facial expression, we treat human facial expression as a distinct modality and incorporate its representation into the language space. Specifically, we extend the vocabulary of the MLLM to include a new token `⟨Expr⟩` that uniquely represents human facial expressions. Given an input text prompt $\mathbf{x}_{txt}$ and/or input image $\mathbf{x}_{img}$, the MLLM f predicts text responses:

$$\hat{\mathbf{y}}_{txt} = f(\mathbf{x}_{img}, \mathbf{x}_{txt}), \tag{1}$$

where $\hat{\mathbf{y}}_{txt} = [t_1, \ldots, t_N]$ is the output sequence of tokens with corresponding hidden states $[h_1, \ldots, h_N]$. When $\mathbf{x}_{txt}$ contains a textual face generation instruction, the predicted response $\hat{\mathbf{y}}_{txt}$ should include a $\langle$`Expr`$\rangle$ token, facilitating further 3D face predictions.

From `⟨Expr⟩` token to 3D facial expression parameters. If one of the output tokens t_n in $\hat{\mathbf{y}}_{\mathbf{txt}}$ is the defined `⟨Expr⟩` token, the model extracts the hidden state as $h^{\langle\texttt{Expr}\rangle} = h_n \in \mathbb{R}^{4096}$ and projects it using the expression decoder η into the latent 3D facial expression parameters $\hat{\psi} = \eta(h^{\langle\texttt{Expr}\rangle}) \in \mathbb{R}^{50}$. The predicted expression parameters $\hat{\psi}$ are combined with a prior head template in the FLAME model [16] to produce a 3D head mesh with diverse expressions. The decoder head η is defined as a lightweight MLP for efficiency and simplicity.

Combining EmoteGPT with downstream applications. As EmoteGPT predicts facial expression parameters based on the FLAME head model, it can be combined with many existing FLAME-based works [28,43,48] as a conditional input, to achieve diverse applications in facial avatar generation. Examples are presented in Figure 2, where EmoteGPT helps produce expressive 3D avatars and expressive 2D face personalized images.

4.3 Training

Training Data Our training data comprises three primary categories:
Text-to-3D expression data. The text-to-3D expression dataset contains paired text descriptions and facial expression parameters, enabling direct supervision for mapping language to expression parameters. We employ the following template to guide the MLLMs during training: `"USER: {description}, can you give the FLAME expression parameters of this person? ASSISTANT: Sure, it is ⟨Expr⟩."` Here, `{description}` denotes a textual expression description and can be replaced with either an explicit or implicit description.

Fig. 3: Qualitative comparison of EmoteGPT with relevant methods. Results are visualized in both the 3DMM space and rendered 3D avatars using GaussianAvatar. Existing models often struggle with complex explicit descriptions and even simple implicit ones. In contrast, EmoteGPT produces semantically faithful and expressive results across diverse inputs. Please zoom in for more details.

Image-to-3D expression data. We collect face images from large-scale pubic face datasets. The face images are processed with SOTA image-based expressive 3D face reconstruction method [4] to obtain the corresponding facial expression parameters. To accommodate the face images into the MLLM training pipeline, we format a template as "`USER: ⟨IMAGE⟩Can you give the FLAME expression parameters of this person? ASSISTANT: Sure, it is ⟨Expr⟩.`", where ⟨`IMAGE`⟩ represents the face image input.

Multimodal Instruction-Following Data. Following LLaVA-1.5 [17], we include general-purpose VQA data to keep model's instruction-following capability.

Overall Training Objective Our overall training objective combines a text-based autoregressive loss with expression reconstruction losses for predicting the ground-truth FLAME expression parameters ψ from either image or text input:

$$\mathcal{L} = \lambda_{\text{txt}}\text{CE}(\mathbf{y}_{\text{txt}}, \hat{\mathbf{y}}_{\text{txt}}) + \lambda_{\text{expr}}|\psi - \hat{\psi}|_2^2 + \lambda_{\text{mesh}}|\mathbf{w} \odot (M(\psi) - M(\hat{\psi}))|_1, \quad (2)$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss between the predicted text tokens $\hat{\mathbf{y}}_{\text{txt}}$ and the ground-truth text tokens $\mathbf{y}_{\text{txt}}$, which helps preserve the model's language understanding and generation capabilities. The expression reconstruction objective consists of two terms: an ℓ_2 loss on the FLAME expression parameters and an ℓ_1 geometric loss on the FLAME mesh. Specifically, $M(\cdot)$ maps FLAME expression parameters to the corresponding 3D mesh vertices, and $\mathbf{w}$ is

a region-dependent vertex weight map applied element-wise via $\odot$. This weighting emphasizes expression-relevant facial regions while reducing the contribution of less relevant head regions. The coefficients λ_{txt}, λ_{expr}, and λ_{mesh} balance the contributions of the text, parameter-space, and mesh-space losses, respectively.

5 Experiments

We evaluate the effectiveness of EmoteGPT in generating 3D facial expressions from natural language prompts. Section 5.1 outlines the experimental setup, including model configurations, datasets, and evaluation metrics. In Section 5.2, we provide quantitative and qualitative comparisons with state-of-the-art methods on both explicit and implicit facial expression generation. Section 5.3 presents an ablation study that examines the impact of different training data modalities. Finally, Section 5.4 demonstrates downstream applications of our method in photorealistic avatar generation and personalized face synthesis.

5.1 Experimental Setup

Network Architecture. Our model is built upon the popular MLLM architecture LLaVA-1.5-7B [17], using Vicuna-v1.5 [44] as the language backbone. To enable efficient fine-tuning, we apply Low-Rank Adaptation (LoRA) [11]. On top of the final layer of the MLLM, we attach a lightweight MLP decoder with GeLU activations [10] and channel dimensions $[5120, 5120, 50]$. This module maps the multimodal representations to FLAME expression parameters, serving as the expression decoder for 3D facial expression prediction.

Implementation Details. We train on 8 NVIDIA A40 GPUs using DeepSpeed [30] with ZeRO [29] optimizer. We use AdamW [20] with a learning rate of 4e$-$5, no weight decay, and a WarmupDecayLR scheduler with 100 warm-up iterations. The loss weights are set to $\lambda_{\text{txt}} = 1.0$, $\lambda_{\text{expr}} = 1.0$, and $\lambda_{\text{mesh}} = 0.01$. The face-region weight $\mathbf{w}$ follows [50]. We use a per-GPU batch size of 8 and 4 gradient accumulation steps. For compatibility with prior FLAME-based methods, we adopt FLAME 2020 model and represent jaw pose with 6D rotations [46].

Datasets. As described in Section 4.3, EmoteGPT is trained using a diverse set of multi-modal data sources: (1) *Text-to-3D expression data:* We use both explicit and implicit textual descriptions from our proposed Txt2Emote dataset as input. The corresponding ground-truth expression parameters are obtained by applying the state-of-the-art expressive face reconstruction method EMOCAv2 [4] to the associated facial images. (2) *Image-to-3D expression data:* To improve the model's ability to interpret real-world expressions, we include facial images from CelebA [19], AffectNet [24], and FFHQ [14], paired with expression parameters extracted using EMOCAv2. (3) *VQA-style multi-modal data:* To preserve the model's general multi-modal reasoning capabilities, we incorporate instruction-following data from LLaVA-v1.5-mix665k [17].

Evaluation Metrics. We evaluate both explicit and implicit text-to-3D facial expression synthesis. Following standard practice in expressive 3D face recon-

Table 3: Quantitative comparisons to state-of-the-art text-to-3D face generation heads. EmoteGPT obtains the best expression synthesis quality compared to other text-driven methods.

Eval	Method	$L1_{Head} \downarrow$	$L1_{Face} \downarrow$	$L1_{Lip} \downarrow$
Exp	HumanNorm	1.03	1.75	4.36
	Portrait3D	0.92	1.51	3.64
	EmoteGPT	**0.62**	**1.01**	**2.37**
Imp	HumanNorm	1.05	1.80	4.64
	Portrait3D	1.03	1.74	4.43
	EmoteGPT	**0.70**	**1.19**	**2.93**

Table 4: User study results comparing EmoteGPT and Portrait3D on explicit and implicit text prompts. '++' indicates strong preference, '+' indicates weak preference. EmoteGPT is preferred in 75% of cases overall, while Portrait3D is preferred in only 14%.

	EmoteGPT		Indifferent	**Portrait3D**	
Prompt Type	++	+		+	++
Explicit	42%	33%	7%	6%	10%
Implicit	46%	30%	10%	10%	2%
Overall	44%	31%	8%	8%	6%

struction [4,31,37], we assess semantic expression fidelity using an emotion recognition network on generated 3D face meshes. This network predicts three types of emotion-related signals from the 3D face mesh: (1) *valence*: positivity or negativity of the expression, (2) *arousal*: intensity of the emotion, and (3) the discrete emotion category. We report Concordance Correlation Coefficient (CCC) for valence and arousal, and top-1 accuracy for emotion classification, using AffectNet annotations as ground truth [24]. For methods that predict FLAME parameters, we further report the L1 distance between predicted and ground-truth 3D point clouds over the head, face and lip regions. Additional metric details are provided in the supplementary.

5.2 Text to 3D Facial Expression Generation

We evaluate our method's ability to generate 3D facial expressions from diverse textual descriptions by comparing it with baselines from two related paradigms: (1) text-to-expression parameter regression methods, which are most closely aligned with our task setting, and (2) recent text-to-3D head generation methods, which also produce expressive 3D faces. Although our approach follows the text-to-expression regression paradigm, we include the latter category to provide a broader comparison.

First, since our method operates within a text-to-expression parameter regression framework, we compare it with baselines in this setting. These include an early text-to-3DMM generation method Describe3D [39], as well as two additional models we design: a CLIP-based regressor and a unimodal LLM-based regressor. These two regressors are trained using the same pipeline like our model to ensure fair and controlled comparison.

We also compare EmoteGPT with state-of-the-art text-to-3D head methods synthesizing full head geometry from text, including HumanNorm [12] and Portrait3D [41]. As these methods typically rely on computationally intensive iterative refinement, large-scale quantitative evaluation is impractical. Instead, we provide qualitative comparisons and conduct a perceptual user study against Portrait3D, the strongest representative in this category.

Quantitative comparison We provide the emotion evaluation results and geometric errors for quantitative comparison on the explicit and implicit-based 3D facial expression synthesis benchmark. To contextualize the difficulty of the benchmark, we also include a static FLAME head with neural expressions as a baseline. As shown in Table 2, the CLIP baseline trained from our Txt2Emote dataset, already surpasses Describe3D in expression generation. Notably, EmoteGPT consistently outperforms both the CLIP and LLM(Vicuna)-based baselines across both explicit and implicit tasks, demonstrating its superior ability at capturing fine-grained and context-dependent expressions. The relatively low performance of the FLAME baseline further indicates that our benchmark comprises diverse and challenging expressions.

Interestingly, Table 2 also highlights the strong generalization ability of our MLLM-based approach. Even when trained without implicit descriptions, EmoteGPT generalizes well to unseen implicit inputs, while maintaining high performance on explicit expressions. This suggests that our architecture and training strategy enable robust expression modeling beyond the training distribution.

We also try to compare EmoteGPT with diffusion-based text-to-3D head generation methods quantitatively. As they are generally time-consuming to synthesis a static head with one single query and it would be impractical to perform large-scale comparison, we subsample 50 explicit and 50 implicit description with balanced emotion distribution in our benchmark for quantitative evaluation. We fit FLAME model for each generated 3D head using Image-based face tracker [4] for a fair comparison.

Table 3 presents the comparison results. Portrait3D performs better than HumanNorm in expressive face synthesis on explicit texts but they perform similarly bad for implicit texts. EmoteGPT consistently obtains the best facial expression synthesis quality than other diffuson-based text-to-3D head methods.

Qualitative comparison. We qualitatively compare EmoteGPT with state-of-the-art text-to-3D face generation methods, including HumanNorm [12] and Portrait3D [41]. As text-to-3D face generation methods synthesize the identity, expression, and texture jointly from texts, to study their expression quality, we fit FLAME model to their generated results and show a comparison in Figure 3.

We observe that all models perform better on explicit prompts, which contain clearer emotion signals, while implicit prompts remain more challenging. HumanNorm, trained with depth- and normal-adapted diffusion on full-body datasets, struggles to interpret facial expressions from natural language inputs. Portrait3D generalizes better due to its strong 3D-aware GAN prior, but still fails to capture fine-grained expression detail—particularly under complex, nuanced descriptions. In contrast, both the CLIP-based baseline and EmoteGPT produce more semantically aligned expressions, with EmoteGPT showing clear advantages in subtle and abstract cases.

Perceptual user study. Evaluating the semantic alignment between textual descriptions and 3D facial expressions remains a challenging task. To assess perceptual quality, we conducted a user study with 25 participants, all with backgrounds in computer graphics. Each participant was shown 20 text prompts

along with rendered expressions from EmoteGPT and Portrait3D. Since Portrait3D outputs full head geometry, we fitted FLAME parameters to its results for a fair comparison. Participants selected the rendering that best matched each prompt. As summarized in Table 4, EmoteGPT was preferred in 76% of cases, while Portrait3D was selected in only 14%, with the remaining 10% indicating no clear preference. These results suggest that EmoteGPT produces facial expressions that more accurately reflect both explicit and implicit textual cues.

Table 5: Ablation study illustrating the impact of different training data modalities on text-to-3D facial expression synthesis. Incorporating facial images data improves EmoteGPT performance on implicit descriptions, demonstrating the effectiveness of multi-modal supervision for enhancing generalization.

Training			Eval							
Text	VQA	Image	Exp	Imp	V-CCC ↑	A-CCC ↑	E-ACC↑	$L1_{Head}$ ↓	$L1_{Face}$ ↓	$L1_{Lip}$ ↓
✓			✓		0.58	0.56	0.52	0.64	1.06	2.49
✓	✓		✓		0.66	0.61	0.56	**0.60**	1.02	2.32
✓	✓	✓	✓		**0.68**	**0.62**	**0.59**	**0.60**	**0.98**	**2.29**
✓				✓	0.55	0.37	0.41	0.71	1.21	2.98
✓	✓			✓	0.55	0.46	0.42	0.67	1.15	2.73
✓	✓	✓		✓	**0.57**	**0.48**	**0.44**	**0.66**	**1.09**	**2.69**

5.3 Ablation experiments

We conduct ablations to evaluate the contributions of different training modalities and the design of the expression grounding token.

Influence of training data As shown in Table 5 and visual examples in supplementary, models trained only on paired text-to-expression data perform worse than those trained with additional multimodal supervision. Incorporating general-purpose VQA data and face image–based expression labels consistently improves performance across input types, demonstrating the effectiveness of our multimodal training strategy.

Design of `<Expr>` token Table 6 compares our dedicated `<Expr>` token with two alternatives: mean pooling over text hidden states and using the `<EOS>` token as the expression representation. Under the same setting, `<Expr>` achieves the best performance across both explicit and implicit inputs. Mean pooling dilutes expression-relevant cues with unrelated textual information, while `<EOS>` entangles expression grounding with sequence termination.

5.4 Further applications

EmoteGPT predicts facial expression within the FLAME model space, making it directly compatible with existing FLAME-based avatar systems. Beyond qualitative comparisons in Figure 3, here we show two downstream applications.

Table 6: Comparison of our ⟨Expr⟩ design with other alternative designs.

Method	Explicit			Impicit		
	$L1_{Head} \downarrow$	$L1_{Face} \downarrow$	$L1_{Lip} \downarrow$	$L1_{Head} \downarrow$	$L1_{Face} \downarrow$	$L1_{Lip} \downarrow$
Mean Pooling	0.62	1.00	2.35	0.68	1.15	2.85
<EOS> Token	0.75	1.21	3.05	0.84	1.43	3.93
<Expr> Token	**0.60**	**0.98**	**2.29**	**0.66**	**1.09**	**2.69**

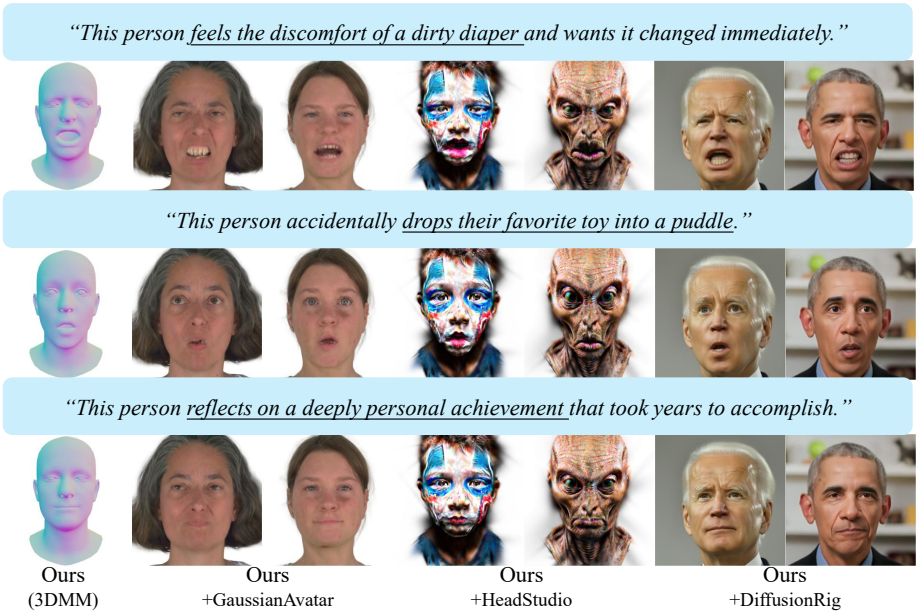

Fig. 4: Diverse applications of EmoteGPT. Combining EmoteGPT with GaussianAvatar, HeadStudio and DiffusionRig enables downstream tasks such as personalized expressive 3D avatar generation and 2D face image synthesis.

– **Expressive 3D avatar generation.** EmoteGPT enables expressive control of 3D avatars through text, complementing existing 3D identity and texture synthesis pipelines like GauaaianAvatar [28] and HeadStudio [48].
– **Expressive 3D-aware personalized image generation.** By combining DiffusionRig [43], EmoteGPT can replace the expression to synthesize identity-preserving, expression-aware face images.

Figure 4 shows the integration of EmoteGPT in diverse pipelines, proving the expressiveness and generality of our approach. More cases are in supplementary.

6 Limitations

Our study adopts FLAME for 3D facial expressions. Although FLAME is widely used and provides a standardized, relatively identity-isolated expression space, it has limited expressiveness. Its training data does not fully cover the diversity of human faces across ages, ethnicities, and cultures, and it may miss subtle geometric details such as wrinkles and fine-grained muscle movements. Moreover, FLAME does not achieve complete identity–expression disentanglement [7]. Therefore, our predicted FLAME parameters should be viewed as an intermediate expression control signal rather than a final high-fidelity facial representation. Such signals can be instantiated by downstream renderers or avatar-generation methods, and our modular framework can be extended to richer blendshape models, neural facial representations, or future 3D morphable models.

Our ground-truth 3D expressions are obtained via a state-of-the-art monocular 3D face estimation method. Although this setup can still struggle under low-quality inputs or extreme poses, it offers a scalable and robust solution for constructing large-scale paired text–3D datasets. We employ AffectNet and EMOCAv2 for their diversity and robustness, and anticipate continued improvements in 3D expression estimation to further enhance our setting.

Finally, our textual supervision relies on captions generated by GPT-4o. While these captions may reflect limited cultural or contextual coverage, GPT-4o remains the strongest vision–language model available for this task. Ongoing advances in large language models are expected to enable richer, fairer, and more inclusive descriptions of facial expressions.

7 Conclusion

We introduced EmoteGPT, a framework for generating 3D facial expressions from natural language using a MLLM with a dedicated `<Expr>` token and a lightweight decoder. To support this task, we introduced Txt2Emote, a dataset of 3D facial expressions paired with rich textual annotations, including explicit appearance-based captions and implicit contextual cues. On benchmark of Txt2Emote, EmoteGPT produces expressions of better quality over existing text-to-3D head and expression methods in emotion accuracy, expression realism, and generalization. Experiments and user studies further validate its ability to generate fine-grained 3D facial expressions, and we demonstrate its versatility in 3D avatar generation, and 2D personalized editing. Overall, this work advances intuitive, text-driven expressive avatar creation and establishes a foundation for broader exploration of MLLMs in 3D human-centric generation.

Acknowledge

Adam Kortylewski acknowledges support via his Emmy Noether Research Group funded by the German Research Foundation (DFG) under Grant No. 468670075.

References

1. Aneja, S., Thies, J., Dai, A., Niessner, M.: Clipface: Text-guided editing of textured 3d morphable models. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings. p. 1–11. SIGGRAPH '23, ACM (Jul 2023). https://doi.org/10.1145/3588432.3591566, http://dx.doi.org/10.1145/3588432.3591566
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999). pp. 187–194. ACM Press (1999)
3. Comas-Massagué, A., Qiu, D., Chai, M., Bühler, M., Raj, A., Gao, R., Xu, Q., Matthews, M., Gotardo, P., Camps, O., Orts-Escolano, S., Beeler, T.: Magicmirror: Fast and high-quality avatar generation with a constrained search space (2024), https://arxiv.org/abs/2404.01296
4. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
5. Daněček, R., Chhatre, K., Tripathi, S., Wen, Y., Black, M., Bolkart, T.: Emotional speech-driven animation with content-emotion disentanglement. ACM (Dec 2023). https://doi.org/10.1145/3610548.3618183, https://emote.is.tue.mpg.de/index.html
6. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (ToG) **39**(5), 1–38 (2020)
7. Egger, B., Sutherland, S., Medin, S.C., Tenenbaum, J.: Identity-Expression Ambiguity in 3D Morphable Face Models . In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–7. IEEE Computer Society, Los Alamitos, CA, USA (Dec 2021). https://doi.org/10.1109/FG52635.2021.9667002, https://doi.ieeecomputersociety.org/10.1109/FG52635.2021.9667002
8. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: Chatting about 3d human pose. In: CVPR (2024)
9. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
10. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR **abs/1606.08415** (2016), http://arxiv.org/abs/1606.08415
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
12. Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., Wang, Q.: Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation (2024)
13. Jiang, Y., Huang, Z., Pan, X., Loy, C.C., Liu, Z.: Talk-to-edit: Fine-grained facial editing via dialog. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

14. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks . IEEE Transactions on Pattern Analysis & Machine Intelligence **43**(12), 4217–4228 (Dec 2021). https://doi.org/10.1109/TPAMI.2020.2970919, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.2970919
15. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR (2024)
16. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813
17. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
21. Ma, Y., Wang, S., Ding, Y., Ma, B., Lv, T., Fan, C., Hu, Z., Deng, Z., Yu, X.: Talkclip: Talking head generation with text-guided expressive speaking styles (2024), https://arxiv.org/abs/2304.00334
22. Manu, P., Srivastava, A., Sharma, A.: Clip-head: Text-guided generation of textured neural parametric 3d head models. In: SIGGRAPH Asia 2023 Technical Communications. SA '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3610543.3626169, https://doi.org/10.1145/3610543.3626169
23. Mendiratta, M., Pan, X., Elgharib, M., Teotia, K., R, M.B., Tewari, A., Golyanik, V., Kortylewski, A., Theobalt, C.: Avatarstudio: Text-driven editing of 3d dynamic human head avatars. ACM Trans. Graph. **42**(6) (dec 2023). https://doi.org/10.1145/3618368, https://doi.org/10.1145/3618368
24. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (Jan 2019). https://doi.org/10.1109/TAFFC.2017.2740923, https://doi.org/10.1109/TAFFC.2017.2740923
25. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu,

K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2024)

26. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
27. Prinzler, M., Zakharov, E., Sklyarova, V., Kabadayi, B., Thies, J.: Joker: Conditional 3d head synthesis with extreme facial expressions (2024), https://arxiv.org/abs/2410.16395
28. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
29. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: memory optimizations toward training trillion parameter models. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC '20, IEEE Press (2020)
30. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3505–3506. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3406703, https://doi.org/10.1145/3394486.3406703
31. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: 3d facial expressions through analysis-by-neural-synthesis (2024),

https://arxiv.org/abs/2404.04104
32. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023)
33. Sun, J., Li, Q., Wang, W., Zhao, J., Sun, Z.: Multi-caption text-to-face synthesis: Dataset and algorithm. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 2290–2298. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475391, https://doi.org/10.1145/3474085.3475391
34. Tan, S., Ji, B., Pan, Y.: Style2talker: High-resolution talking head generation with emotion style and art style. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5079–5087 (2024)
35. Taubner, F., Zhang, R., Tuli, M., Bahmani, S., Lindell, D.B.: MVP4D: Multi-view portrait video diffusion for animatable 4D avatars (2025), https://arxiv.org/abs/2510.12785
36. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5318–5330 (June 2025)
37. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nature Machine Intelligence (2021), https://www.nature.com/articles/s42256-020-00280-0
38. Wei, C., Liu, C., Qiao, S., Zhang, Z., Yuille, A., Yu, J.: De-diffusion makes text a strong cross-modal interface. arXiv preprint arXiv:2311.00618 (2023)
39. Wu, M., Zhu, H., Huang, L., Zhuang, Y., Lu, Y., Cao, X.: High-fidelity 3d face generation from natural language descriptions. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4521–4530. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00439, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00439
40. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
41. Wu, Y., Xu, H., Tang, X., Chen, X., Tang, S., Zhang, Z., Li, C., Jin, X.: Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. ACM Trans. Graph. **43**(4) (Jul 2024). https://doi.org/10.1145/3658162, https://doi.org/10.1145/3658162
42. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
43. Zheng, D., Cecilia, Z., Zhihao, X., Lars, J., Zhuowen, T., Xiuming, Z.: Diffusionrig: Learning personalized priors for facial appearance editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
44. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
45. Zhong, Y., Wei, H., Yang, P., Wang, Z.: Expclip: bridging text and facial expressions via semantic alignment. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Appli-

cations of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024). https://doi.org/10.1609/aaai.v38i7.28594, https://doi.org/10.1609/aaai.v38i7.28594

46. Zhou, Y., Barnes, C., Jingwan, L., Jimei, Y., Hao, L.: On the continuity of rotation representations in neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
47. Zhou, Y., Shimada, N.: 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). pp. 01–08 (2021). https://doi.org/10.1109/FG52635.2021.9666791
48. Zhou, Z., Ma, F., Fan, H., Yang, Z., Yang, Y.: Headstudio: Text to animatable head avatars with 3d gaussian splatting (2024)
49. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
50. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: ECCV (2022)