CLIP-Head: Text-Guided Generation of Textured Neural Parametric 3D Head Models

Anonymous Author(s)
Submission Id: tcom$_1$34

Figure 1: CLIP-Head enables text-driven generation of 3D neural head models in a variety of facial expressions *(left)*, with diverse textures *(middle)* in the form of aligned UV texture maps *(right)*.

ABSTRACT

We propose CLIP-Head, a novel approach towards text-driven neural parametric 3D head model generation. Our method takes simple text prompts in natural language, describing the appearance & facial expressions, and generates 3D neural head avatars with accurate geometry and high-quality texture maps. Unlike existing approaches, which use conventional parametric head models with limited control and expressiveness, we leverage Neural Parametric Head Models (NPHM), offering disjoint latent codes for the disentangled encoding of identities and expressions. To facilitate the text-driven generation, we propose two weakly-supervised mapping networks to map the CLIP's encoding of input text prompt to NPHM's disjoint identity and expression vector. The predicted latent codes are then fed to a pre-trained NPHM network to generate 3D head geometry. Since NPHM mesh doesn't support textures, we propose a novel aligned parametrization technique, followed by text-driven generation of texture maps by leveraging a recently proposed controllable diffusion model for the task of text-to-image synthesis. Our method is capable of generating 3D head meshes with arbitrary appearances and a variety of facial expressions, along with photoreal texture details. We show superior performance with existing state-of-the-art methods, both qualitatively & quantitatively, and demonstrate potentially useful applications of our method.

CCS CONCEPTS

• **Computing methodologies** → **Mesh models**; **Reconstruction**.

SA 2023, Dec 12–15 2023, Sydney, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

CLIP, text-to-3D, parametric models, neural parametric models, UV parametrization.

ACM Reference Format:
Anonymous Author(s). 2023. CLIP-Head: Text-Guided Generation of Textured Neural Parametric 3D Head Models. In *Proceedings of ACM Siggraph Asia (SA 2023)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Faces are a fundamental aspect of communication and identity among humans. In virtual contexts like gaming and mixed reality, the significance of 3D face/head modeling cannot be overstated – it amplifies realism, empowers expressive avatars, aids medical emulators, and influences diverse domains including film, education, and research. Conventional 3D head modeling relies on labor-intensive digital sculpting or intricate 3D scanning, hampering diversity and scalability in digitization. Statistical parametric head models (e.g. FLAME[Tianye Li 2017]), effectively represent shape and expressions of human head in a compact parametric space. These models are though largely successful in achieving 3D head digitization from sparse inputs (e.g. images) and compatibility with standard graphics pipeline, but fall short in capturing intricate surface details due to fixed topology templates, constraining diverse hairstyles. Moreover, the interdependence of their shape and expression parameters restricts controllability [Simon Giebenhain 2023].

In line with contemporary trends in research [Alec Radford 2021; Robin Rombach 2021; Zhang and Agrawala 2023], harnessing natural language for 3D face/head modeling seems not only meaningful but also invaluable. Amidst various proposed methods on these lines [Chi Zhang 2023; Evangelos Ntavelis 2023; Oscar Michel 2021], a recent work ClipFace[Shivangi Aneja 2023] stands out by enabling text-guided generation of photoreal textures and expressions onto 3D morphable head models. The method leverages the parametric geometry of FLAME[Tianye Li 2017], integrated

with an adversarial generative network to synthesize facial appearances in a self-supervised manner. To facilitate text-driven editing and manipulation of texture space, their method employ pre-trained CLIP[Alec Radford 2021]. Another similar work HiFi-Face[Menghua Wu 2023] provides more fine-grain control over geometric manipulation by training a supervised text parser network guided by CLIP encodings to predict descriptive code, which is used to synthesize the initial 3DMM parameters and texture map. However, the text parser requires strong supervision in the form of insanely descriptive text annotations during training. Furthermore, both ClipFace and HiFi-Face suffer from the inherent limitations of parametric models (PCA-based low-dimensional representation lacks disentanglement), constraining the output diversity and flexibility in the generation, while also requiring test-time optimization for texture synthesis.

We posit that an effective 3D head generation method should exhibit specific desired traits such as, its foundational shape and pose representation should be profoundly disentangled, enabling precise manipulation of head geometry and facial expressions with user-friendly text prompts. Additionally, the preference leans towards a single inference step for producing varied appearances and texture styles. On these lines, a recent neural representation Neural parametric head models (NPHM), proposed in [Simon Giebenhain 2023], claims to disentangles the head geometry into two disjoint latent spaces — identity and expression, allowing more granularity in 3D neural head generation by separately decoding sampled latent codes for identity and expressions. However, estimating the values of latent codes for a target identity and expression requires a point cloud representation of the target head and a slow optimization process, limiting the ease and control over generating 3D head geometries.

In this work, we propose a novel method for the text-guided generation of photo-realistic 3D head meshes with varying geometry, expression and high-quality/diverse texture maps. We adapt NPHM representation and facilitate the text-guided estimation of the identity and expression latent codes by introducing two novel mapping networks (MLPs) — one for mapping the CLIP encoded text-prompt embedding to the identity latent code of NPHM, and the other to the expression latent code. Unlike supervised training needed in HiFi-Face[Menghua Wu 2023], we train our networks in a weakly supervised fashion, eliminating the need for pairwise 3D shape/expressions and text description pairs. Utilizing advanced CLIP-based diffusion models, we aim to employ Latent Diffusion[Robin Rombach 2021] for text-guided high-quality texture synthesis across diverse 3D head geometries. To ensure harmony between the text-driven synthesis and 3D head geometry, we propose a custom-trained ControlNet[Zhang and Agrawala 2023] architecture, where we feed UV normal maps of the 3D head mesh as the control hint for fine-grain steering of UV texture (RGB) map generation. However, the 3D scan meshes exhibit varying and unstructured UV parametrization, demanding semantically meaningful and aligned UV maps as valuable ControlNet hints. To achieve this, we introduce a technique to roughly align 3D head mesh UV coordinates, enabling the localization of similar regions within the same UV space. During inference, the aligned UV normal map of the generated NPHM head mesh is fed into ControlNet, synthesizing UV texture maps guided by simplistic text prompts. In summary, our key contributions are as follows:

- *A novel method to provide text-guided control over a neural head representation for highly controllable 3D head generation.*
- *A novel technique to automatically align the UV texture coordinates of NPHM meshes, enabling text-guided high-fidelity texture synthesis in a single feed-forward pass.*

2 METHOD

Given an input text prompt, first the *Geometry Synthesis* module generates an NPHM head mesh in accordance with the textual description. Subsequently, this mesh is fed to *Aligned UV Parametrization* module for the seam estimation and yielding a coherent UV map with projected surface normals. Finally, the *Texture Synthesis* module generates a UV texture map, guided by the input text prompt.

In regard to neural head representation adopted from NPHM [Simon Giebenhain 2023], let the disentangled latent vectors for face identity and expression (as in NPHM) be z_{id} and z_{exp}, respectively. In order to decode z_{id} and z_{exp}, the NPHM employ two pre-trained MLP-based decoders (trained on a dataset of head scans), $\mathcal{F}_{id}$ & $\mathcal{F}_{exp}$, yielding a neural SDF representation of the head [see Figure 2 *(right)*]. Subsequently, a polygonal head mesh $\mathcal{M}$ is extracted by performing χ, consisting of two operations — SDF query and marching cubes. Thus, the entire process of NPHM is given as,

$$\mathcal{M} = \chi(\mathcal{F}_{id}(z_{id}) + \mathcal{F}_{exp}(z_{id}, z_{exp})) \quad (1)$$

where"+" is the deformation of the neural field described in [Simon Giebenhain 2023].

2.1 Geometry Synthesis

This module aims to map a region of the CLIP's embedding space to the NPHM's latent space. We achieve this in a novel fashion by proposing an identity mapping network MLP_{id} and an expression mapping network MLP_{exp}, which learns to map the CLIP embedding vector to the corresponding latent vectors z_{id} & z_{exp}, respectively. Starting with a text prompt, CLIP's text encoder generates ψ, then z_{id} and z_{exp} are obtained via MLP_{id} and MLP_{exp} respectively. These latents are then used in Equation 1 to yield the head mesh $\mathcal{M}$, as illustrated in Figure 2.

We first discuss the training strategy of identity mapping network MLP_{id}, which requires corresponding ground truth latent vectors z_{id} associated with specific CLIP embedding vectors ψ. In the absence of an annotated dataset, we propose a novel automated approach to generate such pairs using the ControlNet method. As shown in Figure 2(right), starting with randomly sampled z_{id} from NPHM's identity latent space, a neutral expression head mesh $\mathcal{M}_{neutral}$ is generated, serving as the base. This mesh is then randomly rotated about a vertical axis and rendered to acquire normal map I_{norm}. Employing the ControlNet, guided by a template prompt and control hint I_{norm}, we synthesize images resembling facial features and geometry captured by I_{norm}. Each resulting image's embedding vector ψ is used with its corresponding z_{id} as paired samples for MLP_{id} training. Approximately $10k$ such pairs are generated for this purpose. This approach leverages weak supervision, capturing facial shape information implicitly through image

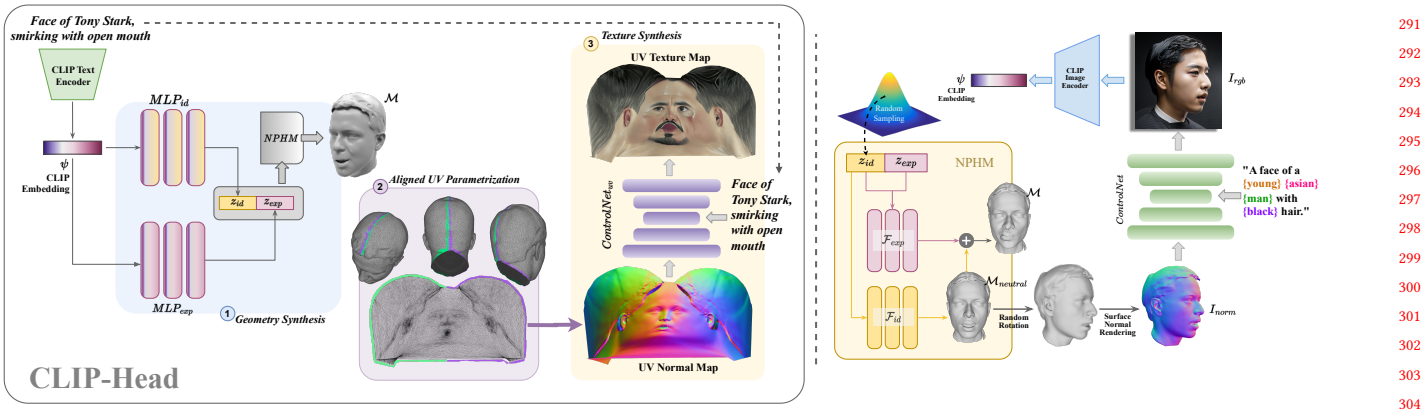

Figure 2: Pipeline of the proposed framework *(left)*. Data generation procedure for training MLP_{id} *(right)*.

decoding. During inference, by using a text prompt to generate ψ through CLIP's text encoder, followed by mapping through MLP_{id}, an accurate identity latent vector z_{id} can be obtained due to CLIP's text-image mapping properties.

For training expression mapping network MLP_{exp}, we propose a different strategy for generating training pairs of z_{exp}'s and the associated ψ's. We leverage the dataset proposed in [Simon Giebenhain 2023], where 23 common expressions per subject are captured. Since we have different mapping networks for facial geometry (appearance) and expressions, we propose to select a single identity and use its expressions for generating the training pairs. To do so, we first provide a label for each of the 23 expressions in the NPHM dataset, e.g. *happy, sad, angry, pouting, laughing*, etc. Given latent vector z_{exp} associated with an expression, we then curate random prompts using the template — **"*A face of a* {young/middle-age/old} {ethnicity} {man/woman/person} *with* {color} *hair,* {expression}."**. These prompts are encoded by the CLIP's text encoder, generating corresponding ψ's which are used as input, along with the corresponding z_{exp} for training MLP_{exp}. Once both the mapping networks are trained, given a text prompt's CLIP encoding ψ, MLP_{id} predicts z_{id} and MLP_{exp} predicts z_{exp}. Both the predicted latent vectors are then fed to Equation 1 to obtain head mesh $\mathcal{M}$.

2.2 Aligned UV Parametrization

This module performs UV parametrization of arbitrary head meshes such that similar parts of all the head meshes lie approximately in the same region in the UV space, as shown in Figure 2. This rough alignment is required for training the *Texture Synthesis* module, which we explain later in subsection 2.3. To obtain a low-distortion UV parametrization of the given head mesh $\mathcal{M}$, we first estimate a seam to cut the mesh. The back part of the head seems a natural choice to avoid artifacts on the face. Since, all the NPHM head meshes obtained via Equation 1 fall in the same coordinate system, we identify the common bisecting plane to designate the seam. Thereafter, the boundary vertices of the head mesh (seam vertices + neck boundary vertices) are mapped to a fixed 2D curve in the UV space (depicted by green/purple color-coded curve in Figure 2). Since the boundary of the mesh is fixed, we now compute two harmonic functions — one for U and one for the V coordinate. Each harmonic function uses the fixed vertices on the curve as boundary constraints to estimate a harmonic parametrization [Eck, 2005] for the remaining vertices.

2.3 Texture Synthesis

In the final stage of the proposed framework, we propose to synthesize a coherent, high-quality UV texture map for the head mesh $\mathcal{M}$. Since we are interested in generating text-driven UV texture maps (which are essentially 2D RGB images), employing latent-diffusion-powered ControlNet for this task seems an apt choice for generating diverse textures. However, the task is not trivial. We propose to train a network $ControlNet_{uv}$ on our own, which takes some form of control hint and generates a text-guided UV texture map for $\mathcal{M}$. Since the control hint should somehow describe the UV layout of $\mathcal{M}$ so that the generated texture map follows the same UV layout, we propose to project the face normals onto the UV space to obtain a UV normal map to use as control hint (as shown in Figure 2). Now, the motivation for aligning the UV layout of arbitrary head meshes becomes more evident, as the control hint should have semantically meaningful characteristics which can be interpreted easily by $ControlNet_{uv}$. For training, we employ the same aligned UV parametrization (proposed in the previous section) for the head scans in the NPHM dataset to obtain aligned UV normal map and UV texture map pairs. Note that the default unaligned UV layouts of the head scans are unstructured. For automatically generating accurate text prompts to be used during training, we render the head scans and pass the rendered image to BLIP-2 for captioning. Once trained, the proposed $ControlNet_{uv}$ is capable of generating a high-quality UV texture map for a given head mesh $\mathcal{M}$, controlled by the underlying UV normal map and guided by the text prompt given by the user. Optionally, the generated texture map can be enhanced by applying super-resolution using [Xintao Wang 2021]. All the quantitative and qualitative evaluations are done without this enhancement.

3 RESULTS & EVALUATION

Qualitative Evaluation: We demonstrate qualitative results of the proposed framework, where we show 3D head generation on a wide variety of styles (Figure 3(a)) and styles (Figure 3(b)). Figure 4 shows a qualitative comparison with SOTA methods, where our

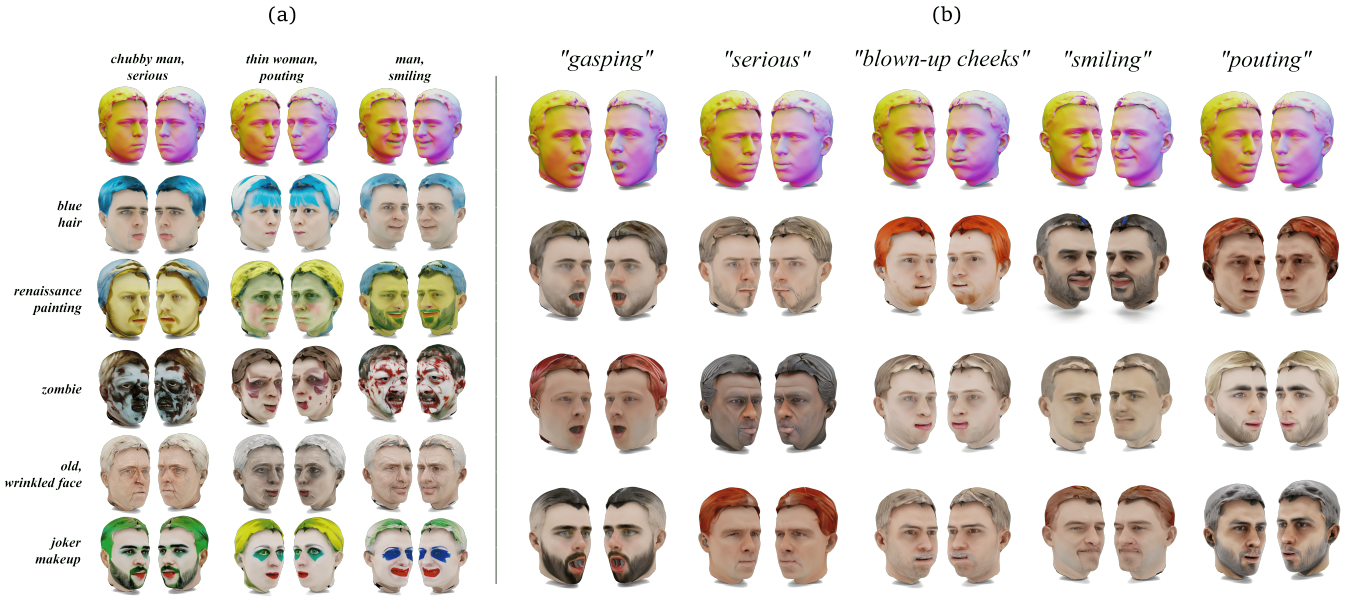

Figure 3: Qualitative Results: (a) Text-driven generation and stylization of 3D head meshes. (b) Text-driven generation of 3D head meshes with varying expressions.

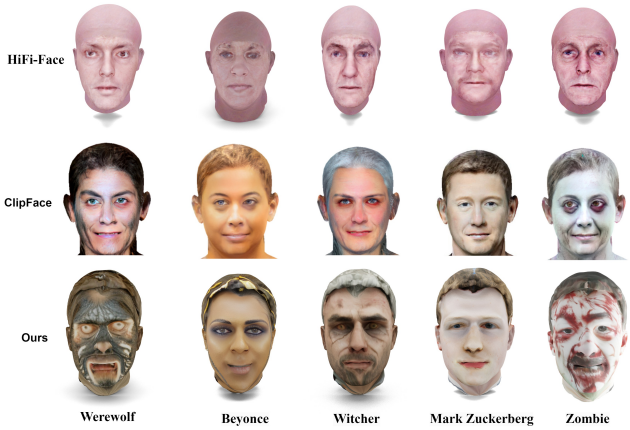

Figure 4: Qualitative comparison with existing SOTAs.

Method	CLIP Score ↑		
	ViT-H/14	ViT-L/14	ViT-B/16
ClipFace	0.287 ± 0.041	0.289 ± 0.039	0.307 ± 0.023
HiFi-Face	0.229 ± 0.033	0.236 ± 0.031	0.300 ± 0.018
CLIP-Head	**0.292 ± 0.035**	**0.303 ± 0.039**	**0.315 ± 0.021**

Table 1: Quantitative comparison with SOTA methods.

generation results seem more convincing and true to the input text prompt (as revealed in the user study).

Quantitative Evaluation: We compare the CLIP Score[Shivangi Aneja 2023] (metric explained in supplementary) of our method with existing SOTAs in Table 1, where we outperform other methods in all three variants of the encoders. This further solidifies the claim that we have a higher similarity between the input text and rendered mesh image.

Please refer to the supplementary for ***user study, extended qualitative/quantitative evaluation, discussion, limitations and 360° video renderings.***

REFERENCES

Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*. PMLR.

Yijun Fu Zhenglin Zhou Gang Yu Zhibin Wang Bin Fu Tao Chen Guosheng Lin Chunhua Shen Chi Zhang, Yiwen Chen. 2023. StyleAvatar3D: Leveraging Image-Text Diffusion Models for High-Fidelity 3D Avatar Generation. arXiv:2305.19012 [cs.CV]

Kyle Olszewski Chaoyang Wang Luc Van Gool Sergey Tulyakov Evangelos Ntavelis, Aliaksandr Siarohin. 2023. AutoDecoding Latent 3D Diffusion Models. arXiv:2307.05445 [cs.CV]

Linjia Huang Yiyu Zhuang Yuanxun Lu Xun Cao Menghua Wu, Hao ZhuB. 2023. High-fidelity 3D Face Generation from Natural Language Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Richard Liu Sagie Benaim Rana Hanocka Oscar Michel, Roi Bar-On. 2021. Text2Mesh: Text-Driven Neural Stylization for Meshes. *arXiv preprint arXiv:2112.03221* (2021).

Dominik Lorenz Patrick Esser Bjorn Ommer Robin Rombach, Andreas Blattmann. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR* abs/2112.10752 (2021). arXiv:2112.10752

Angela Dai Matthias Niessner Shivangi Aneja, Justus Thies. 2023. ClipFace: Text-Guided Editing of Textured 3D Morphable Models. In *ACM SIGGRAPH 2023 Conference Proceedings* (Los Angeles, CA, USA) *(SIGGRAPH '23)*.

Markos Georgopoulos Martin Runz Lourdes Agapito Matthias Nießner Simon Giebenhain, Tobias Kirschstein. 2023. Learning Neural Parametric Head Models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Michael J. Black Hao Li Javier Romero Tianye Li, Timo Bolkart. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017).

Chao Dong Ying Shan Xintao Wang, Liangbin Xie. 2021. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]