AnimateMe: 4D Facial Expressions via Diffusion Models

Dimitrios Gerogiannis, Foivos Paraperas Papantoniou, Rolandos Alexandros Potamias, Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, and Stefanos Zafeiriou

Imperial College London, UK
{d.gerogiannis22, f.paraperas, r.potamias, a.lattas, s.moschoglou, s.ploumpis, s.zafeiriou}@imperial.ac.uk

Abstract. The field of photorealistic 3D avatar reconstruction and generation has garnered significant attention in recent years; however, animating such avatars remains challenging. Recent advances in diffusion models have notably enhanced the capabilities of generative models in 2D animation. In this work, we directly utilize these models within the 3D domain to achieve controllable and high-fidelity 4D facial animation. By integrating the strengths of diffusion processes and geometric deep learning, we employ Graph Neural Networks (GNNs) as denoising diffusion models in a novel approach, formulating the diffusion process directly on the mesh space and enabling the generation of 3D facial expressions. This facilitates the generation of facial deformations through a mesh-diffusion-based model. Additionally, to ensure temporal coherence in our animations, we propose a consistent noise sampling method. Under a series of both quantitative and qualitative experiments, we showcase that the proposed method outperforms prior work in 4D facial expression synthesis by generating high-fidelity extreme expressions. Furthermore, we applied our method to textured 4D facial expression generation, implementing a straightforward extension that involves training on a large-scale textured 4D facial expression database.

Keywords: 4D Expression · Diffusion Models · Graph Neural Networks

1 Introduction

In the field of computer graphics and human-computer interaction, 3D face modeling [4, 5, 7, 38, 39, 51] and animation are becoming increasingly crucial. With the evolution of digital interactions, there is a rising demand for realistic 3D avatars that can generate emotions with high fidelity. While controllable and dynamic 2D facial expression generation is well-studied [9, 18, 32, 52, 55, 57], its 3D counterpart remains unexplored due to its inherent complexity. This lack of research emphasizes the need for advancements in 4D facial expression generation. While much of the 3D face animation domain focuses on speech animation [1, 13, 19, 25, 31, 35–37, 42, 47, 49, 50, 53, 58], research on expression animation [33, 40] is limited. Although diffusion models have been sparingly applied to

facial speech animation techniques [1,31,35,47,49], the results have shown considerable promise. In contrast, these models have been extensively and successfully employed in the field of 3D human motion, demonstrating their effectiveness and versatility [2, 11, 14, 44, 48, 60]. Despite the proven success of diffusion models in 3D animation, their application in 4D facial expression remains unexplored. This gap in the literature inspired us to develop our method, aiming to investigate their potential in this domain.

However, typical diffusion model architectures are challenging to apply directly to 3D structures. Using meshes as our 3D face representation, we introduce the first 3D diffusion process tailored to operate directly within the mesh space, enabling the application of diffusion models for 4D facial expression generation. Our method achieves this by employing Graph Neural Networks (GNNs) as denoising diffusion models. This novel approach paves the way for broader utilization of GNNs in diffusion processes for mesh generation. Moreover, regarding temporal coherence in animations, our method diverges from the traditional diffusion models methods. Conventional video approaches [6, 21, 22, 24, 45, 61] typically handle temporal dependencies via architectural adjustments incorporating temporal modules, generating frames collectively. In contrast, inspired by ideas presented in [26, 29, 56], our approach modifies the traditional DDPM algorithm by introducing a straightforward yet effective sampling strategy tailored to our specific challenge, termed consistent noise sampling. This intuitive sampling strategy not only solidifies temporal coherence but also significantly improves the generation time.

Our method capitalizes on diffusion models, presenting a marked divergence from the competition. Although some methods have been explored for 4D facial expression generation [33, 40], they often fall short in producing high-fidelity extreme expressions, while our method successfully accomplishes it. This advantage is primarily attributed to the diffusion models' capability to handle complex distributions, such as the extreme deformations accompanying intense facial expressions. The strength of our model is further enhanced by utilizing the entirety of the mesh space. This offers superior capturing capabilities compared to traditional blendshapes [10] and landmarks [33].

In summary, our work offers the following key contributions:

- The first, to the best of our knowledge, diffusion process formulation operating directly on the mesh space with GNNs proposed as denoising models.
- The first, to the best of our knowledge, fully data-driven approach to customizable 4D facial expressions, utilizing diffusion models.
- A dynamic diffusion models sampling strategy for 3D facial animation, that is extended to both geometry and texture generation.

2 Related Work

2.1 3D Facial Animation Generation

Since the introduction of the seminal 3DMMs [4, 5], multiple methods have been proposed to model facial animation using expression blendshapes [10, 12].

Nonetheless, they train on static 3D meshes, and can only rely on unrealistic linear interpolation to represent 3D facial motion.

Several methods have tried to tackle the limitations of 3D facial motion synthesis by utilizing either audio or speech features. In an early attempt to model speech features along with facial motion, the authors of [25] presented a subject-specific model to produce facial motion from audio but faced limitations in adapting to varied speakers. A similar method was proposed in [13] where a static neutral mesh of a given identity was fed to the network for animation based on speech input features showcasing flexibility across diverse speakers. Several other approaches were built on top, that utilize RNNs along with facial action units [37] and LSTMs [53]. Following studies significantly improved the realism of 3D speech animation by disentangling emotion from speech [36, 42], while Transformer-based approaches [19,50,58] have demonstrated promising results as well. Only very recently, a few works have focused on animating a 3D facial mesh via a diffusion process coupled with a voice input signal [1,31,35,47,49]. Most of these works deal with denoising and generating animations based on input speech that is later mapped into blendshape expression parameters [35, 47]. Only [49] works directly on the mesh domain utilizing a motion decoder. Even though this line of work is promising for facial animation based on speech input, none of the aforementioned methods deal with labeled (guided) expression generation.

Guided 4D expression generation is an understudied problem with only a few methods able to achieve satisfactory results. The work proposed in [40] first attempted guided 4D expression generation on the 4DFAB dataset [12], combining LSTMs with GNNs. It notably addressed the challenge of animating extreme expressions for the first time. While this work utilizes a GNN architecture as a decoder, similar to our method, our approach differentiates itself by adopting a mesh diffusion process, thereby enhancing the expression capturing capabilities. Similar to this research direction, the authors in [33] proposed a two-stage framework employing a sparse-to-dense decoder mapping sparse 3D facial landmark displacements to dense ones for motion generation, complemented by a GAN architecture for landmark sequence creation driven by expression labels. Although effective, this method falls short of capturing extreme expressions and detailed facial features as efficiently as our approach, as demonstrated in our experiments section. This discrepancy is likely due to the superiority of diffusion models over GANs [15], and our comprehensive use of the entire mesh rather than just landmarks.

2.2 Diffusion Models for Animation

Since the introduction of diffusion models in the 2D image domain [23, 46], the computer vision literature witnessed staggering advancements in image and video synthesis when compared to previous GAN-based approaches [15, 43]. Regarding 2D animation, [24] was the first work to extend diffusion models to video synthesis by adapting the traditional denoising architecture to accommodate 3D data. Following this pioneering work, many studies focused on maintaining temporal coherence in animation, while achieving high resolution and frame

rate through a variety of methods. These include integrating temporal modules, refining denoising architectures for video data, implementing cascaded super-resolution techniques, extending text-to-image models to videos, extending the latent diffusion model paradigm to video diffusion models, or even combinations of these approaches, often leveraging the strengths of each to achieve superior outcomes [6, 21, 22, 45, 61]. The unique approaches of [29] and [26] diverge from the typical approaches to handling temporal coherence. The authors of [29] innovated by treating video frames as non-independent instances by refining the DDPM paradigm to introduce a shared base noise and a time-variant residual noise across frames. Similarly, [26] departs from traditional random sampling by enforcing motion dynamics between the latent codes. Together, these two approaches emphasize the benefits of using consistent noise patterns across frames serving as inspiration for our work.

Extending diffusion models to generate animation in the 3D domain is an ambitious task given the vast amounts of 3D data required over time and the inherently greater complexity of 3D structures compared to images. Recent efforts have leveraged these models to generate motion for 3D models. Apart from the few diffusion-based 3D speech animation methods referenced in the previous subsection, the vast majority of such endeavors are focused on generating 3D human motion, which is a well-studied research area. Numerous diffusion based methods ranging from body motion [2, 11, 14, 44, 48, 60] to hand gestures motions [3, 59], have emerged in the literature. While most of these methods are text-conditioned, others utilize different mechanisms such as 3D landmarks [16] and 3D object points [27]. Nevertheless, none of the aforementioned methods tackles the problem of labeled expression animation which is an understudied problem due to the scarcity of 4D facial expression data, unlike speech and human motion data.

2.3 Diffusion Models on the 3D Space

Although 2D diffusion models have been extensively explored and understood, particularly in terms of denoising model architectures and their applications, the exploration in the 3D domain lags behind because of the complexities involved in 3D modeling. Only a handful of studies have explored diffusion processes directly on 3D structures, and these primarily work on point clouds. The pioneering work of [28] adapted traditional diffusion models for point clouds introducing a probabilistic approach rooted in thermodynamic diffusion processes. Similarly, another approach [62] combined diffusion models with point-voxel representations for shape generation. Recent methods have followed more sophisticated approaches by operating on the 3D latent space [30, 54] taking inspiration from [43]. The last two methods can also generate meshes, but they achieve this by reconstructing surfaces from the generated point clouds and lack detail [30, 54]. To the best of our knowledge, no existing diffusion model directly operates on mesh points while preserving their inherent connectivity to generate meshes.

3 Method

Our method learns to animate a mesh of neutral expression towards a target expression, guided by a signal that indicates both the progression and intensity of the resulting expression animation, enabling extensive customization.

To implement the animation mechanism, our framework introduces a novel mesh diffusion process tailored for fixed topology meshes. The frames of each animation used for training are processed by expressing them as deformations from the neutral mesh. Interestingly, while our method is dynamic, our diffusion model is trained in a conventional static manner. A key feature is our intuitive consistent noise sampling strategy, designed specifically for our problem. This not only ensures temporal coherence, resulting in smooth animations but also accelerates the generation process. Fig. 1 provides an overview of our frame generation method, while Fig. 2 depicts our consistent noise sampling strategy. In the following subsections, we detail each component and explain the innovations of our method.

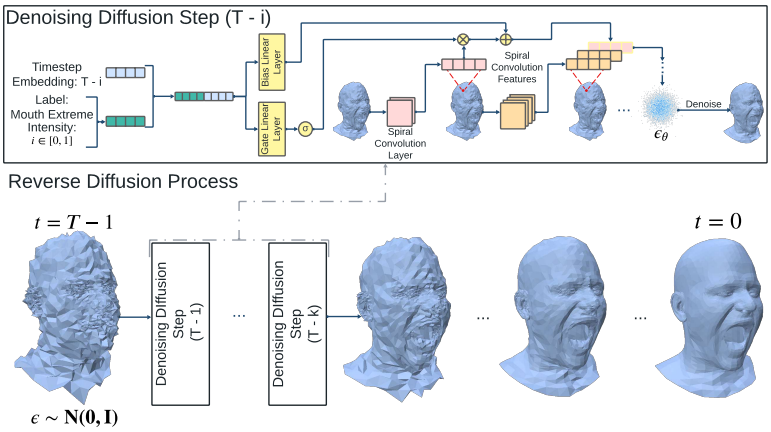

Fig. 1: Overview of the proposed frame generation method: Our method generates frames by integrating a point cloud DDPM with an SCN denoising model, conditioned on a concatenated expression and timestep conditioning. It employs spiral convolutional layers, modulating output features with a simple gating and bias mechanism tailored to the conditions. Throughout this process, noise is predicted and systematically subtracted at each timestep until the frame is completely denoised and thus generated. While the method operates on deformations, for visualization, we apply them to the neutral mesh for all timesteps, to show the temporal evolution of the diffusion process.

3.1 Formulating Diffusion Processes on the Mesh Space

Our mesh diffusion formulation builds upon [28]. The core idea of this work is to tailor the diffusion process specifically for point clouds. The objective is to

reconstruct a point cloud $\mathbf{X}^{(0)}$ of a desired shape, defined by a shape latent $\mathbf{z} \in \mathbb{R}^{N\times 1}$. This latent is derived from an encoder $\mathbf{z} = E_\phi(\mathbf{X}^{(0)})$, and the reverse diffusion process is conditioned on it. Starting from pure noise, the method progresses to generate the point cloud.

Consistent with traditional DDPMs, the training objective is to maximize the data's log-likelihood, $\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{X}^{(0)})]$. By leveraging the variational lower bound (ELBO) and further derivations, a simplified training objective is formulated, akin to the approach presented in [23]:

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{N} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{a}_t}\mathbf{x_i}^{(0)} + \sqrt{1-\bar{a}_t}\boldsymbol{\epsilon}, t, E_\phi(\mathbf{X}^{(0)})) \right\|^2 \tag{1}$$

Fixed topology meshes are the dominant representation of human faces [17]. Hence, our objective is to adapt the point cloud diffusion process to operate directly on such meshes. However, challenges arise when examining the original point cloud diffusion process for mesh generation. The primary issue arises from the integration of the loss function, which is focused on denoising at a point level, with the denoising model $\boldsymbol{\epsilon_\theta}(\mathbf{x_i}^{(t)}, t, \mathbf{z})$, characterized by its point cloud MLP architecture. The latter treats point clouds as unordered sets due to their inherent permutation invariance. Recognizing that meshes essentially represent strongly structured point clouds defined by their connectivity, such a property becomes a problem, since order and structure are crucial for them. Moreover, point cloud MLPs initially process individual points independently and only aggregate this data in subsequent layers, leading to challenges in accurately capturing the finer local details found in meshes.

To facilitate mesh generation, we employ GNNs as denoising diffusion models in a novel manner. This adaptation leads to a structure-aware, sophisticated diffusion process that not only effectively denoises, but also preserves the connectivity. More specifically, we replace the point cloud MLP denoising model with a Spiral Convolutional Network (SCN) [8, 20]. This replacement is denoted by $\boldsymbol{s_\theta}(\mathbf{x_i}^{(t)}, t, \mathbf{c})$ where $\mathbf{c}$ represents the conditioning of the model, serving a role similar to the shape latent $\mathbf{z}$ in the original method.

3.2 Training in Line with Static Diffusion Models

Typically, dynamic diffusion models that generate modalities, such as videos, train on the entire sequence of animation frames, operating in a whole-animation fashion. In contrast, our approach to training is one frame at a time. Specifically, for a given animation $\mathbf{X} = \{\mathbf{x_i} : i = 0, \ldots, K-1\}$, we select a single frame $\mathbf{x_i} \in \mathbb{R}^{N\times 3}$. We then extract the expression stage information for this frame, denoted as $\mathbf{e_i} \in \mathbb{R}^{1\times M}$, where M represents the number of possible expression categories. This vector $\mathbf{e_i}$ encodes both the category and intensity of the expression, derived from the animation's expression signal $\mathbf{E} = \{\mathbf{e_i} : i = 0, \ldots, K-1\}$. Additionally, we obtain the neutral mesh associated with the animation, denoted as $\mathbf{x_0}$. We use it to express the current frame $\mathbf{x_i}$ as deformations relative to it, yielding $\mathbf{d_i} = \mathbf{x_i} - \mathbf{x_0}$ with $\mathbf{d_i} \in \mathbb{R}^{N\times 3}$. From there, training follows the paradigm of

static diffusion models with the denoising model conditioned on the expression stage and timestep, using the following loss function:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\| \boldsymbol{\epsilon} - \boldsymbol{s}_\theta(\sqrt{\bar{a}_t}\mathbf{d_i}^{(0)} + \sqrt{1-\bar{a}_t}\boldsymbol{\epsilon}, t, \mathbf{e_i}) \right\|^2 \tag{2}$$

A significant advantage of our approach is its computational efficiency. Traditional methods require loading entire animation sequences for training, which is computationally demanding, especially with large meshes. Our frame-by-frame approach avoids this, enabling high-resolution mesh training without performance limitations.

3.3 Using the Entire Mesh versus Relying on Landmarks

We train on the entire mesh, capturing the complex dynamics of facial expressions, including the finest details, drawing inspiration from [40]. This holistic approach, utilizing the comprehensive deformation mesh $\mathbf{d_i} = \mathbf{x_i} - \mathbf{x_0}$ for each frame i, surpasses both traditional blendshape-based techniques [10,12] and landmarks based methods [33]. Traditional techniques often fail to represent extreme deformations due to their linear limitations. In contrast, while landmark-based methods capture a significant portion of facial motions through a set of landmarks, they may overlook the fine details of facial dynamics. It is worth noting that even though the methodologies presented in [33] yield mesh displacements, their core training still relies on sparse landmarks.

3.4 Consistent Noise Sampling

While the conditioning of the diffusion process on the expression stage might impose some sort of temporal coherence, it is by no means enough for the smoothness required for expression animation. To bridge this gap, building on the ideas presented in [26, 29, 56], we propose a modification of the original DDPM sampling algorithm for our problem, named consistent noise sampling, rooted in two primary observations.

Firstly, within the diffusion process, noise drives sample diversity. However, in facial expression animation, maintaining temporal coherence requires careful management of this diversity, as it can otherwise hinder it. To tackle this issue, we propose employing a consistent noise strategy across all animation frames to ensure smooth transitions, acknowledging the minimal differences typically present between consecutive frames. This approach involves applying consistent noise both at the start of denoising and throughout the following denoising steps, ensuring that each frame within an animation maintains coherence. By sampling and maintaining a consistent initial noise implementation $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and noise sequence for denoising $\mathbf{z} = \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t = T-1, \ldots, 1$ with $\mathbf{z}_0 = \mathbf{0}$ across frames, our model effectively differentiates its output based on the specific expression stage information, ensuring smooth animations.

Secondly, in diffusion models, generation progresses step-wise in a hierarchical manner: earlier timesteps address broader structures, while later steps refine details, such as expression dynamics. Therefore, our method applies the full range of denoising steps $t = T-1, \ldots, 0$ for the generation of the first frame of an animation and then reduces steps for the following frames. To implement our approach, we first apply the full range of denoising steps, $t = T-1, \ldots, 0$, for generating the initial frame of an animation. Once we reach a late-stage denoised version at timestep t_s for the first frame, denoted as $\hat{\mathbf{d}}_{\mathbf{0}}^{\mathbf{t_s}}$, we then initiate the generation of subsequent frames from this advanced denoised state utilizing only the remaining range of timesteps $t = t_s, \ldots, 0$. This means that the initial noise prediction for the first frame is given by $\hat{\epsilon_{\mathbf{0}}}^{\mathbf{T-1}} = \boldsymbol{s_\theta}(\epsilon, T-1, \mathbf{e_0})$. For subsequent frames, however, the initial noise prediction adjusts to $\hat{\epsilon_{\mathbf{i}}}^{\mathbf{t_s}} = \boldsymbol{s_\theta}(\hat{\mathbf{d}}_{\mathbf{0}}^{\mathbf{t_s}}, t_s, \mathbf{e_i})$, with t_s effectively serving as the new starting point ($T'-1 = t_s$) for these frames. This strategic adjustment allows for a more efficient generation process by reducing the number of required timesteps while maintaining high fidelity.

To generate diverse animations for the same expression with our sampling strategy, one simply samples different noise implementations and applies them consistently across all frames. A final advantage of our sampling, which might not be evident at the start, is that once the initial frame is generated, all subsequent frames can be produced concurrently, further improving generation speed.

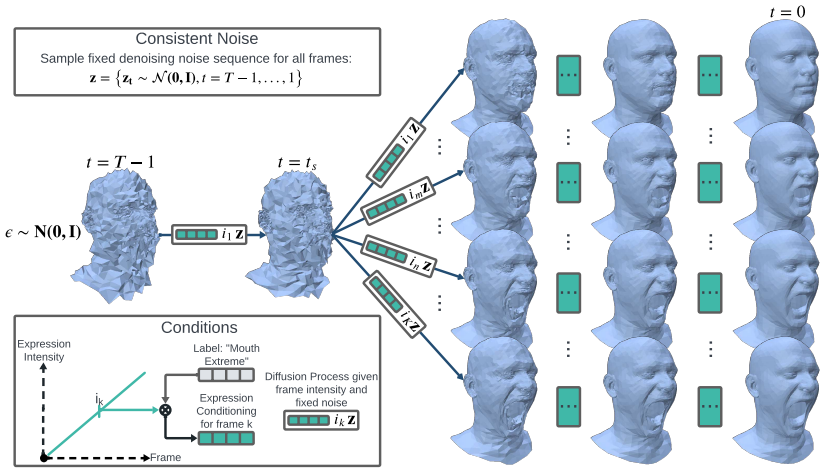

Fig. 2: Animation generation via consistent noise sampling: The process initiates by sampling the initial noise ϵ and the denoising noise sequence $\mathbf{z}$ over $T-1$ timesteps. The diffusion process begins with the first frame using the full range of timesteps. Upon reaching a late denoised stage at t_s, the generation for subsequent frames starts in parallel from this denoised state, utilizing only the remaining t_s timesteps. All frames share the same denoising sequence $\mathbf{z}$, with differences arising from the expression intensity.

4 Experiments

To demonstrate the superiority of our approach over existing methods, we undertook both quantitative and qualitative experiments focused on 4D facial expression animation using the CoMA [41] dataset. This dataset features 12 unique identities, each with 12 varied expressions that are extreme and diverse, making it an ideal benchmark for evaluating 4D facial expression generation methods.

Additionally, we applied our approach to the generation of textured 4D facial expressions by training on the MimicMe [34] dataset, a large-scale database that provides both geometry and textures. We followed the same training protocol used with CoMA for the geometric data and employed a texture animation method using a latent diffusion model. This allowed us to effectively combine the geometry with texture sequences, producing realistic textured 4D facial expressions.

4.1 4D Facial Expression Evaluation

Preprocessing. To ensure meaningful comparisons and address the limitations of [33] which is limited to a fixed frame count per animation, all methods are standardized to produce 40 frames. Following the logic of [33], we select subsequences that transition from a neutral to an extreme expression and using interpolation or selection, we ensure consistent length for all. Our preprocessing then quantifies the expression progression by calculating deformations from the neutral mesh for each frame of an animation and smoothing it appropriately. Furthermore, we use global scaling for intensity via an animation extremeness factor, which is necessary for customizable approaches such as ours and [40] to grasp expression intensity information. This factor is normalized across animations of the same expression, providing a consistent framework of assessing expression intensity. Our preprocessing pipeline, designed to equip the model with both progression and intensity information is versatile enough to be applied to any dataset. For more details, we refer the reader to the supplementary material.

Training Setup. To accurately evaluate the generalization of our method, we opted to split the CoMA expression animations subject-wise, excluding all animations of the test subjects from training. For the experiments, we follow the settings outlined in [33] and modify the approach described in [40] to accommodate the CoMA dataset. Our diffusion model's noise schedule uses 1000 steps, starting from $t_1 = 1e-4$ to $t_T = 0.02$. The late denoising strategy is configured with $t_s = 400$, meaning that for subsequent frames beyond the first, we employ only 400 timesteps, initiating from the late denoised version of the first frame, 400 steps before concluding the diffusion process. Our model was trained for 5600 epochs with a batch size of 32, using the Adam optimizer with an initial learning rate of 0.001 and a learning rate scheduler to finally reduce to $1e-4$.

Quantitative Evaluation. Given that our training animations transition from neutral to extreme states, we expect [33] to inherently learn to generate extreme expressions. To ensure fairness with [33], which isn't designed for customizable intensity levels, we demonstrate the adaptability of our model. We generate expression animations at maximum intensity through our global intensity scaling strategy (Ours-Extreme) and, to mitigate bias, also produce animations across a spectrum of intensities (Ours-Varying), highlighting our method's adaptability. Further enhancing comparison fairness, we apply local intensity scaling (Ours-Local), training our model exclusively on expression progression values. This approach simulates a non-customizable framework, closely mirroring [33], even though it limits the potential of our method. Considering the customizable aspect and significantly worse relative performance of [40], we pair it with global scaling to achieve the best results, ensuring a truly fair comparison between all methods.

To quantitatively assess the generated expression animations across all methods, we employ the standard metrics. Expression classification is a key benchmark for evaluating 4D facial expression methods. Combining the classifier solutions presented in [33, 40], we adopt a similar approach. Our classifier involves a two-stage process. Initially, we use Principal Component Analysis (PCA) to capture the core variations of facial expressions by encoding deformations from the neutral state. This encoding is applied to all animations used to train the models, resulting in a PCA encoder that effectively reduces the dimensionality of the meshes while preserving their expressive spatial features. Following this, we deploy an LSTM-based classifier that operates on these PCA-encoded animations. By processing the temporal sequence of PCA-encoded mesh deformations, our LSTM model effectively captures the dynamic nature of facial expressions. The LSTM output is then sequentially fed through two fully connected layers, with the final layer responsible for class prediction. The second metric utilized for evaluation is the specificity measure, defined as the per-frame average Euclidean distance between the generated animations and the ground truth. This metric serves as an estimate of how closely the generated animations resemble the actual ones, thus acting as a direct indicator of generation quality. Additionally, we employ the Fréchet Inception Distance (FID) metric to evaluate the quality of the generations. For our method, we employ this metric only for the version that mimics the non-customizable framework (Ours-Local) because this provides the fairest comparison for FID. When combined with classification accuracy and specificity measure, these metrics together provide a comprehensive framework for the quantitative analysis of 4D facial expressions.

For all subsequent experiments involving our method and MO3DGAN [33], we generate 50 animations per subject and expression due to their stochastic nature. In contrast, for the deterministic LSTM approach [40], only a single animation is produced for each subject and expression.

As can be seen in Tab. 1, our method surpasses the other two methods across all metrics, demonstrating its superiority. Remarkably, it achieves a similar classification performance with the ground truth, underscoring the quality of our generated expressions. In more detail, the specificity error increases for the final

Table 1: Classification accuracy, average specificity error (mm) and FID.

Method	Accuracy (%)	Specificity (mm)	FID
GT	77.77	0	29.15
MO3DGAN [33]	67.22	2.18	52.21
LSTM [40]	45.12	3.2	59.14
Ours-Extreme	75.94	1.78	—
Ours-Local	72.88	1.74	43.01
Ours-Varying	78.90	1.61	—

frames, as shown in Fig. 3, because these frames correspond to the most extreme expressions. Despite this, every version of our method consistently shows lower error, particularly in these challenging later frames, highlighting our method's capability to accurately generate complex and extreme expressions. To further reinforce our argument about our method's effectiveness in capturing extreme expressions, we include heatmaps in Fig. 4, clearly showcasing our method's preeminence.

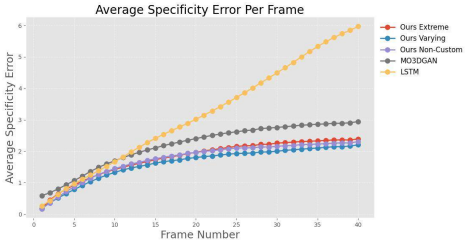

Fig. 3: Average per frame specificity error (mm) between the proposed and the baseline methods LSTM [40] and MO3DGAN [33].

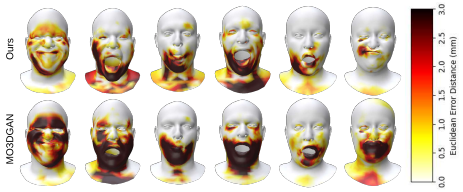

Fig. 4: Comparison of final frame generations with their respective ground truths, between ours and MO3DGAN [33]. LSTM [40] results are omitted for brevity and due to significantly worse performance.

Qualitative Evaluation. Beyond quantitative analysis, we offer qualitative results for a deeper, subjective evaluation. Due to space constraints, we qualitatively compare with the best-performing method. Fig. 5 showcases generations of extreme expression animations from both methods for a visual comparison. Our method excels by producing more extreme and expressive animations, highlighting our capability to generate high-fidelity expressions. Additionally, as illustrated in Fig. 6, our method produces animations of high smoothness. Finally, our model excels in creating highly diverse expression animations, with Fig. 7 showcasing examples from the expression categories with the greatest inherent diversity.

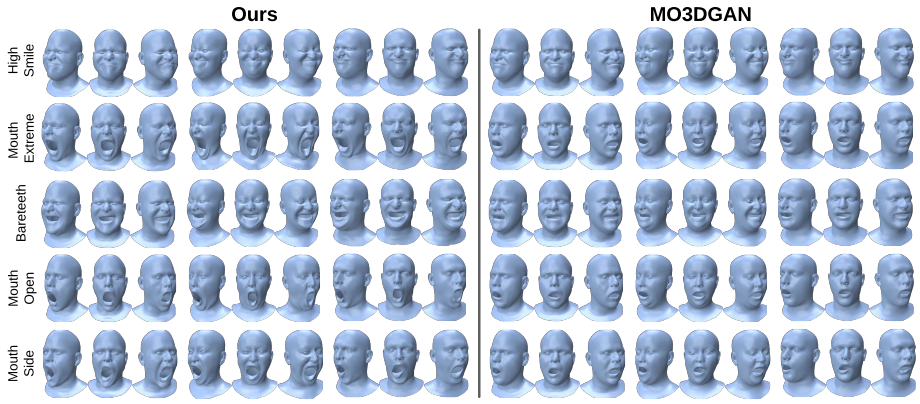

Fig. 5: Qualitative comparison of extreme expression generations between ours and MO3DGAN [33]. Final expressions are illustrated in the main paper. For full-length dynamic 4D expressions, please refer to the supplementary material.

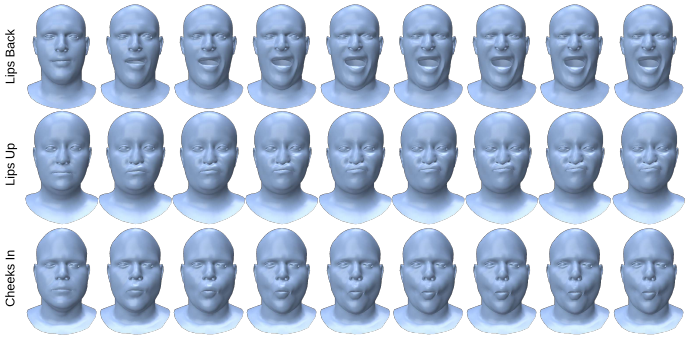

Fig. 6: Progression of 4D expressions for three different identities and expressions.

Fig. 7: Diversity of generated expressions.

4.2 Textured 4D Animation on Large Scale Datasets

In our final experiment, we implemented a simple extension of our method to generate textured 4D facial expressions using the diverse MimicMe [34] dataset, which includes 4,700 subjects, each performing the same expressions. Despite the dataset's variety, it poses challenges such as low frame rates and non-uniform expression initiation points, with some animations transitioning directly between expressions without starting from a neutral state. To overcome these issues, we manually annotated and curated a subset of subjects and their expression animations. By interpolating between frames, we increased the frame count for both texture and geometry. Consequently, we created a refined dataset of 345 subjects, each demonstrating the six fundamental expressions with 40 frames per expression, ensuring a close match between geometry and texture.

Building on the principles of our geometry diffusion model, we implemented a latent diffusion model (LDM) [43] to generate sequences of textures. This approach enhances the standard LDM architecture by conditioning it on the neutral latent (i.e., the neutral texture encoded using the LDM's encoder), expression intensity, and label of the generated frame. The conditioning mechanisms include channel-wise concatenation for the neutral latent and cross-attention for the expression intensity and label signals. To ensure a meaningful and representative latent space, we trained the autoencoder component of the LDM with all textures from the MimicMe [34] dataset.

Our training methodology mirrors that of the geometry model, focusing on one frame at a time. This parallel training process ensures a unified learning strategy across both geometry and texture models. For inference, we employ the same consistent noise sampling strategy as used in the geometry model, which has also proven effective for animating textures in our context. Additionally, by utilizing a dataset significantly larger than CoMA's [41] 12 subjects, we have equipped the geometry diffusion model with identity information. This is achieved by encoding the neutral mesh using a spiral convolutional encoder [8,20] that is trained jointly with the diffusion model. The resulting encoding, concatenated with the other conditioning inputs, guides the reverse diffusion process, effectively integrating identity into the generative process.

The resulting framework generates textured 4D facial expressions given expression information, a neutral mesh and texture. It achieves this through the application of the geometry model to generate the geometry of the frames and the texture model to create the corresponding textures for each frame. Qualitative results of our method are showcased in Fig. 8 for subjective evaluation.

Fig. 8: Generated textured 4d facial expressions using our framework. Notably, while our framework doesn't generate texture on top of geometry, it consistently produces texture sequences that qualitatively align with the corresponding geometries. Expressions progress from left (neutral) to right (apex).

5 Conclusion

In this work, we introduce AnimateMe, a novel diffusion-based model for fully customizable 4D facial expression generation. Leveraging our novel mesh diffusion process with the GNN serving as the denoising model, we facilitate expression generations of high fidelity, significantly surpassing the existing state-of-the-art. Paired with our consistent noise sampling strategy, our model ensures the production of smooth animation sequences. We demonstrated the adaptability of our model by extending it to textured animation. This extension signifies our method's potential for application in large-scale databases, offering a unified framework for both geometry and texture modeling. AnimateMe introduces a 4D method for modeling extreme expressions, effectively addressing a challenge that has not been solved as effectively as possible in the existing literature. To the best of our knowledge, it also proposes the first diffusion model to utilize a GNN as a denoising model.

Acknowledgements

S. Zafeiriou and part of the research were funded by the EPSRC Fellowship DEFORM (EP/S010203/1) and EPSRC Project GNOMON (EP/X011364/1).

References

1. Aneja, S., Thies, J., Dai, A., Nießner, M.: Facetalk: Audio-driven motion diffusion for neural parametric head models (2023)
2. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: Large-scale text-conditional 3d human motion generation. arXiv preprint arXiv:2305.09662 (2023)
3. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: Neural sign actors: A diffusion model for 3d sign language production from text. arXiv preprint arXiv:2312.02702 (2023)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. Seminal Graphics Papers: Pushing the Boundaries, Volume 2 (1999), https://api.semanticscholar.org/CorpusID:203705211
5. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence **25**(9), 1063–1074 (2003)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
7. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7213–7222 (2019)
8. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
9. Bouzid, H., Ballihi, L.: Facial expression video generation based-on spatio-temporal convolutional gan: Fev-gan. Intelligent Systems with Applications **16**, 200139 (Nov 2022). https://doi.org/10.1016/j.iswa.2022.200139, http://dx.doi.org/10.1016/j.iswa.2022.200139
10. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2014). https://doi.org/10.1109/TVCG.2013.249
11. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
12. Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
13. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019)

14. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9760–9770 (2023)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
16. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023)
17. 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)
18. Fan, L., Huang, W., Gan, C., Huang, J., Gong, B.: Controllable image-to-video translation: a case study on facial expression generation. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19, AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33013510, https://doi.org/10.1609/aaai.v33i01.33013510
19. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers (2022)
20. Gong, S., Chen, L., Bronstein, M., Zafeiriou, S.: Spiralnet++: A fast and highly efficient mesh convolution operator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
21. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2023)
22. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models (2022)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
25. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
26. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
27. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
28. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2837–2845 (2021)
29. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation (2023)
30. Lyu, Z., Wang, J., An, Y., Zhang, Y., Lin, D., Dai, B.: Controllable mesh generation through sparse latent point diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 271–280 (2023)

31. Ma, Z., Zhu, X., Qi, G., Qian, C., Zhang, Z., Lei, Z.: Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. arXiv preprint arXiv:2402.05712 (2024)
32. Otberdout, N., Daoudi, M., Kacem, A., Ballihi, L., Berretti, S.: Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 848–863 (2022). https://doi.org/10.1109/TPAMI.2020.3002500
33. Otberdout, N., Ferrari, C., Daoudi, M., Berretti, S., Bimbo, A.D.: Sparse to dense dynamic 3d facial expression generation (2022)
34. Papaioannou, A., Gecer, B., Cheng, S., Chrysos, G., Deng, J., Fotiadou, E., Kampouris, C., Kollias, D., Moschoglou, S., Songsri-In, K., et al.: Mimicme: A large scale diverse 4d database for facial expression analysis. In: European Conference on Computer Vision. pp. 467–484 (2022)
35. Park, I., Cho, J.: Said: Speech-driven blendshape facial animation with diffusion. arXiv preprint arXiv:2401.08655 (2023)
36. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20687–20697 (2023)
37. Pham, H.X., Cheung, S., Pavlovic, V.: Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 80–88 (2017)
38. Ploumpis, S., Ververas, E., O'Sullivan, E., Moschoglou, S., Wang, H., Pears, N., Smith, W.A., Gecer, B., Zafeiriou, S.: Towards a complete 3d morphable model of the human head. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4142–4160 (2020)
39. Ploumpis, S., Wang, H., Pears, N., Smith, W.A., Zafeiriou, S.: Combining 3d morphable models: A large scale face-and-head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10934–10943 (2019)
40. Potamias, R.A., Zheng, J., Ploumpis, S., Bouritsas, G., Ververas, E., Zafeiriou, S.: Learning to generate customized dynamic 3d facial expressions (2020)
41. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European conference on computer vision (ECCV). pp. 704–720 (2018)
42. Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1173–1182 (October 2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
45. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=nJfylDvgzlq

46. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
47. Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. pp. 1–11 (2023)
48. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
49. Thambiraja, B., Aliakbarian, S., Cosker, D., Thies, J.: 3diface: Diffusion-based speech-driven 3d facial animation and editing. arXiv preprint arXiv:2312.00870 (2023)
50. Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized speech-driven 3d facial animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20621–20631 (October 2023)
51. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7346–7355 (2018)
52. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation (2017)
53. Tzirakis, P., Papaioannou, A., Lattas, A., Tarasiou, M., Schuller, B., Zafeiriou, S.: Synthesising 3d facial motion from "in-the-wild" speech. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 265–272 (2020)
54. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems **35**, 10021–10039 (2022)
55. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3an: Disentangling appearance and motion for video generation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5263–5272 (2020). https://doi.org/10.1109/CVPR42600.2020.00531
56. Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7378–7387 (2023)
57. Wu, X., Zhang, Q., Wu, Y., Wang, H., Li, S., Sun, L., Li, X.: F^3a-gan: Facial flow for face animation with generative adversarial networks. IEEE Transactions on Image Processing **30**, 8658–8670 (2021). https://doi.org/10.1109/tip.2021.3112059, http://dx.doi.org/10.1109/TIP.2021.3112059
58. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)
59. Zhang, F., Ji, N., Gao, F., Li, Y.: Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In: International Conference on Multimedia Modeling. pp. 231–242. Springer (2023)
60. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
61. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models (2023)

62. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)