# Collaborative Tooth Motion Diffusion Model in Digital Orthodontics

**Yeying Fan[1], Guangshun Wei[2], Chen Wang[1], Shaojie Zhuang[1], Wenping Wang[2,3], Yuanfeng Zhou[1]***

[1]School of Software, Shandong University, China
[2]Department of Computer Science, The University of Hong Kong, China
[3]Texas A&M University, USA
fyy@mail.sdu.edu.cn, guangshunwei@gmail.com, chen.wang@mail.sdu.edu.cn, shjie.zhuang@gmail.com, wenping@tamu.edu, yfzhou@sdu.edu.cn

## Abstract

Tooth motion generation is an essential task in digital orthodontic treatment for precise and quick dental healthcare, which aims to generate the whole intermediate tooth motion process given the initial pathological and target ideal tooth alignments. Most prior works for multi-agent motion planning problems usually result in complex solutions. Moreover, the occlusal relationship between upper and lower teeth is often overlooked. In this paper, we propose a collaborative tooth motion diffusion model. The critical insight is to remodel the problem as a diffusion process. In this sense, we model the whole tooth motion distribution with a diffusion model and transform the planning problem into a sampling process from this distribution. We design a tooth latent representation to provide accurate conditional guides consisting of two key components: the tooth frame represents the position and posture, and the tooth latent shape code represents the geometric morphology. Subsequently, we present a collaborative diffusion model to learn the multi-tooth motion distribution based on inter-tooth and occlusal constraints, which are implemented by graph structure and new loss functions, respectively. Extensive qualitative and quantitative experiments demonstrate the superiority of our framework in the application of orthodontics compared with state-of-the-art methods.

## Introduction

The tooth motion generation is a key task in digital orthodontics, which significantly assists dentists in efficiently making diagnoses or treatment plans. Specifically, given the initial pathological and target ideal tooth alignments, we need to generate the motion process that transforms the tooth alignment from the initial to the target state, as shown in Fig. 1. At the same time, the tooth motion generation needs to meet the requirements of no collision, reasonable interaction among multi-tooth, shortest and smooth transition paths, simultaneously. The whole process of tooth motion generation is time-consuming and laborious, and the quality of tooth motion generation relies heavily on the subjective experience of dentists and technicians. Therefore, developing an automatic and data-driven tooth motion generation method is essential. However, existing data-driven digital orthodontics methods mainly focus on the basic tooth data processing task (Cui et al. 2021; Qiu et al. 2022; Ma et al. 2020; Cui et al. 2022) and the tooth alignment target prediction (Li et al. 2020a; Wei et al. 2020; Wang et al. 2022). Due to its complexity and challenge, the data-driven tooth motion generation method remains a blank area.

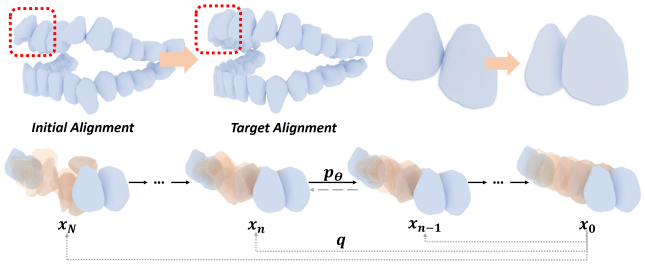


Figure 1: Our goal is to generate a tooth motion process of the intermediate tooth alignments (orange) given the initial alignment (blue) and the target alignment (blue). It shows the sampling process from the predicted motion distribution is first from a random normal distribution (leftmost column) and gradually denoised using a learned denoiser into the final predictions (rightmost column).

Previous approaches formulate tooth motion generation task as a multi-agent motion planning problem by using traditional path planning algorithms such as A-Star (Li and Yang 2011), particle swarm (Ma et al. 2021), genetic (Li, Li, and Li 2009) and artificial bee colony algorithms (Li et al. 2020b), which require solving complex optimizations and separate treatment of the upper and lower jaw, resulting in the ignorance of the occlusal relationship. Some data-driven tooth alignment target prediction methods (Wei et al. 2020; Wang et al. 2022) generate the tooth motion process in an iterative manner. However, these methods directly employ tooth point cloud data with numerous parameters, leading to limited tooth motion steps and error accumulation. Instead, we remodel the tooth motion generation problem as a distribution fitting problem, which samples the motion from the learned distribution with the given initial and target alignments, as shown in Fig. 1. We aim to directly model the motion distribution of the entire process.

Diffusion-based generative models have achieved remarkable success in path planning (Janner et al. 2022), human

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

motion generation (Li, Liu, and Wu 2023; Tevet et al. 2022; Chen et al. 2023; Dabral et al. 2023), and trajectory prediction (Gu et al. 2022; Jiang et al. 2023). They demonstrate the benefits of the diffusion model, including flexible behavior synthesis, long-horizon scalability, and variable-length plans in the motion generation and planning task (Janner et al. 2022). These advantages are well-suited for fitting motion distributions. However, several challenges still exist when applying diffusion models to fit the tooth motion distribution. First, fitting tooth motion distribution requires a simple and precise condition to guide the training and sampling process. This condition needs to accurately describe tooth position and posture while simultaneously characterizing the different types of teeth. Second, the tooth motion process involves collaboration between individual teeth. It is challenging to fit a tooth motion distribution that satisfies multi-tooth collaboration and orthodontic medical requirements.

In light of these challenges, we propose a novel method that synthesizes the tooth motion process via a conditional diffusion model. Firstly, we propose a tooth latent representation method as a condition of the diffusion model, containing the tooth frame and the shape latent code, which reduces network parameters compared to directly predicting the tooth motion process using point clouds. The tooth frame is constructed from tooth landmarks and axes (Wei et al. 2022; Yf et al. 2022), as shown in Fig. 2. The tooth frame contains position and posture information. We also need to characterize the shape of different types of teeth, which is essential for multi-tooth interaction, so we obtain the tooth shape latent code through a tooth point cloud auto-encoder. Secondly, we propose a collaborative tooth motion diffusion model. The diffusion model aims to fit the tooth motion distribution, and the multi-tooth collaboration is reflected in the feature level and loss functions. Since the topological relationship among teeth is fixed, we introduce a multi-tooth interaction module using the graph neural network to achieve motion information interaction between teeth at the feature level during the motion process. Meanwhile, we define a set of loss functions as soft restrictions to help the fitted tooth motion distribution obtain the most realistic results and conform to orthodontics requirements as much as possible.

The main contributions of this work are:

- We propose the first diffusion-based framework for tooth motion generation under the given initial and target tooth alignment states, which remodels the problem as a conditional diffusion process. Our method enables concurrent generation of all timesteps.
- We introduce a tooth latent representation method that includes the tooth frame, representing the position and posture of the tooth model, and the tooth latent shape code, representing the geometric information. This approach enables the provision of more precise conditional guidance.
- We design multiple constraints, the multi-tooth interaction graph module and new loss functions, to provide effective collaborative multi-tooth and supervision for tooth motion generation while complying with orthodontic medical requirements.

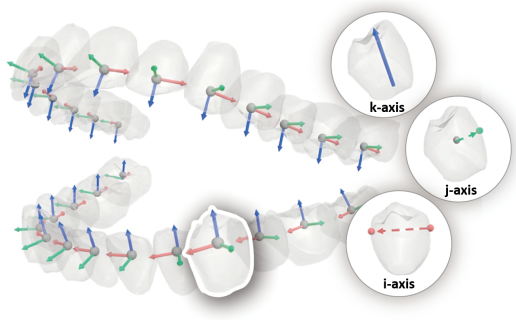


Figure 2: Tooth frame representation. The gray, green, and red dots represent the centroid, facial axis point (FA), and contact point (CO). The red, green, and blue arrows represent the i, j, and k (tooth long axis) axes.

## Related Works

**Tooth Feature Representation.** Deep learning-based dental tasks using the intraoral scan model include tooth segmentation (Cui et al. 2021; Qiu et al. 2022; Cui et al. 2022), tooth classification (Ma et al. 2020), tooth landmark/axis detection (Wei et al. 2022; Yf et al. 2022), tooth alignment target prediction (Wei et al. 2020; Yang et al. 2020; Wang et al. 2022), and so on (Song et al. 2021; Zhang et al. 2022). Most of them first take the segmented point cloud of an intraoral scan model as input and then utilize the point cloud feature extraction network (Qi et al. 2017a,b; Wu, Qi, and Fuxin 2019; Wang et al. 2019) to extract tooth point cloud features for downstream works. Although the extracted dental feature, in the high dimension level, contains tooth geometric shape features, it does not represent the position and posture information of the tooth concretely. In this paper, we propose a latent representation for teeth, enabling our network to better focus on position, posture, and characterize shape information.

**Tooth and Human Motion Synthesis.** Human motion synthesis is a fundamental task in computer animation (Tevet et al. 2022). It aims to generate human motion through the human skeleton. Motion in-betweening is derived from the motion prediction problem, where the resulting motion is constrained on some given past and future keyframes (Qin, Zheng, and Zhou 2022). Most works rely on a two-branch, including autoregressive-based and non-autoregressive-based. The autoregressive-based methods (Harvey et al. 2020; Tang et al. 2022) predicted the intermediate motions step by step. The non-autoregressive-based methods (Kaufmann et al. 2020; Qin, Zheng, and Zhou 2022; Kim et al. 2022) can generate the whole intermediate motions simultaneously. While human and tooth motion generation shares similarities, there are notable application gaps in the following areas: (a) Current human motion generation methods focus on single (Tevet et al. 2022; Raab et al. 2023) or several persons (Guo et al. 2022; Xu, Wang, and Gui 2023; Shafir et al. 2023). (b) Significant distinctions exist in the representation of human and tooth characteristics. To fill this gap, we propose a multi-tooth collaborative diffusion framework for tooth motion generation.

**Motion Diffusion Models.** Diffusion Generative Models have succeeded dramatically in many research areas, including image generation (Nichol et al. 2022; Rombach et al. 2022), 3D Vision (Lyu et al. 2022), video modeling (Ho et al. 2022), and medical imaging (Wyatt et al. 2022). Besides, diffusion models also achieved competitive results in planning and motion generation tasks, including path planning (Janner et al. 2022), trajectory prediction (Gu et al. 2022; Jiang et al. 2023), human motion generation (Li, Liu, and Wu 2023; Tevet et al. 2022; Chen et al. 2023), and procedure planning in instructional videos (Wang et al. 2023). Instead of step-by-step prediction, it can generate all timesteps of planning, trajectory, and motion. Unlike the human motion generation method (MDM) (Tevet et al. 2022) regarding motion in-betweening as an image inpainting problem (Kaufmann et al. 2020) by unconditional diffusion model and only considering a single agent, we utilize conditional diffusion to generate the whole tooth motion process for multi-tooth.

# Methodology

In this section, we present the details of our multi-tooth collaborative diffusion model for tooth motion generation. We first introduce the setup for this problem and overview. Then, we introduce the framework which contains two parts: the tooth latent representation and the multi-tooth collaboration diffusion model. Finally, we show the design of multiple constraints. An overview of the collaborative tooth motion diffusion model is illustrated in Fig. 3.

## Problem Notation and Formulation

Each tooth alignment $\mathcal{T}$ contains three types of information $\mathcal{T} = \{P_v, F_v, \mathcal{Z}_v | P_v \subseteq \mathbb{R}^{M\times 3}, F_v = \{\boldsymbol{o}; \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}, \mathcal{Z}_v \subseteq \mathbb{R}^z, v \in \mathcal{V}\}$, where $\mathcal{V}$ denotes the set of tooth labels. $P_v$ denotes the point cloud of the $v$-th tooth, which contains $M$ points in $\mathbb{R}^3$. $F_v$ is the $v$-th tooth frame, which consists of four elements: $\boldsymbol{o}$, $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$ denote the centroid, $\boldsymbol{i}$-axis, $\boldsymbol{j}$-axis, and $\boldsymbol{k}$-axis by a three-dimension vector respectively. $\mathcal{Z}_v$ represents the tooth shape latent code. A complete tooth alignment comprises 28 teeth. The tooth labels are assigned as $\mathcal{V} = \{11-17, 21-27, 31-37, 41-47\}$ according to the Federation Dentaire Internationale teeth representation.

Given the initial tooth alignment $\mathcal{T}_{initial}$ and the target alignment $\mathcal{T}_{target}$, our goal is to plan a motion process $\boldsymbol{X}$ to transform the tooth alignment from $\mathcal{T}_{initial}$ to $\mathcal{T}_{target}$, eventually obtaining all an intermediate tooth motion process. The $\boldsymbol{X} = \{\boldsymbol{x}_v^{1:L} | v \in \mathcal{V}\}$ represents a set of single tooth motion $\boldsymbol{x}_v$ with the motion process length $L$ and $\boldsymbol{x}_v^l \in \mathbb{R}^6$ denotes a 6 DoF (6 degrees of freedom) transformation parameter $(r_x, r_y, r_z, t_x, t_y, t_z)$. The core idea is to train a diffusion model to learn the tooth motion distribution by the proposed tooth latent representation as the condition, and then sample from the learned motion distribution at the inference phase.

First, the tooth latent representation aims to encode the position and posture of given tooth alignments by using the tooth frame $F$ and shape latent code $\mathcal{Z}$. Then, the diffusion model aims to approximate a multi-tooth motion distribution $p(\boldsymbol{x}^{1:L})$ for realizing the transformation from the $\mathcal{T}_{initial}$ state to $\mathcal{T}_{target}$ state.

## Tooth Latent Representation

To address the complex tooth motion generation problem, we propose a tooth latent representation to provide more precise conditional guidance both for the training and sampling process. This method consists of two encoders: one for encoding the tooth frame to accurately describe the tooth position and posture, and the other for encoding tooth shape to capture the diverse geometric morphology of the tooth model. The former focuses on the global feature of the teeth during the motion process, while the latter focuses on the local feature acting on the collaboration between multiple teeth. This approach avoids the computational complexity of using tooth point cloud or mesh data directly.

**Tooth Frame Encoder.** To provide a more precise description of tooth position and posture, we propose a tooth frame encoder similar to TAD-Net (Yf et al. 2022) which is a point cloud-based learning method for automatic tooth axes detection. Unlike various tooth axes that are mutually independent, the three axes of the tooth frame exhibit pairwise orthogonality. To meet this characteristic, we treat three axes in the tooth frame as a whole and encode it as a quaternion. Then, we employ a point-wise rotation transformation prediction network based on point cloud learning to abstract the tooth frame from the tooth point cloud. In other words, the tooth frame $F = \mathcal{E}_{frame}(P)$ can be considered the skeleton of the tooth model.

**Tooth Shape Encoder.** The tooth shape information is an essential factor affecting multi-tooth motion generation. Since the tooth geometric morphology is implied in the tooth point cloud data, we propose a tooth shape encoder to abstract complex tooth shape information from the tooth point cloud $P$. In particular, we use an AutoEncoder to construct the tooth point cloud reconstruction network, which consists of a PointNet encoder $\mathcal{E}_z$ and a decoder $\mathcal{D}$ with a multi-layer perceptron (MLP) to learn a representative and low-dimensional latent space $\mathcal{Z} = \mathcal{E}_z(P)$ for diverse teeth.

## Multi-tooth Collaborative Diffusion

It is challenging to generate tooth motions while considering the collaboration among teeth. We propose a multi-tooth collaborative diffusion model to overcome this. Firstly, the model learns the motion distribution of each tooth through a single-tooth conditional diffusion model. Subsequently, the multi-tooth interaction module enables inter-tooth motion interaction at the feature level.

**Single Tooth Conditional Diffusion Models.** Diffusion probabilistic models (Ho, Jain, and Abbeel 2020) as a kind of generative model that can gradually anneal the noise from a Gaussian distribution to a data distribution $p(\boldsymbol{x})$ by learning the noise prediction from a $N$ length Markov noising process, given $\{\boldsymbol{x}_n\}_{n=1}^N$. For tooth motion generation, we introduce the single tooth conditional diffuser $p_\theta(\boldsymbol{x}_{0,v}^{1:L} | c_v)$ for each tooth $\mathcal{T}_v$ to generate tooth motion $\hat{\boldsymbol{x}}_{0,v}^{1:L}$ through iterative denoising. $\boldsymbol{x}_{0,v}^{1:L}$ is drawn from the data distribution of the $v$-th tooth motion transformation. The condition $c$ of each tooth diffuser contains four parts: the initial $F_{initial}$ and target $F_{target}$ tooth frames represent the tooth motion initial and target states, and the offset $F_{offset}$ information

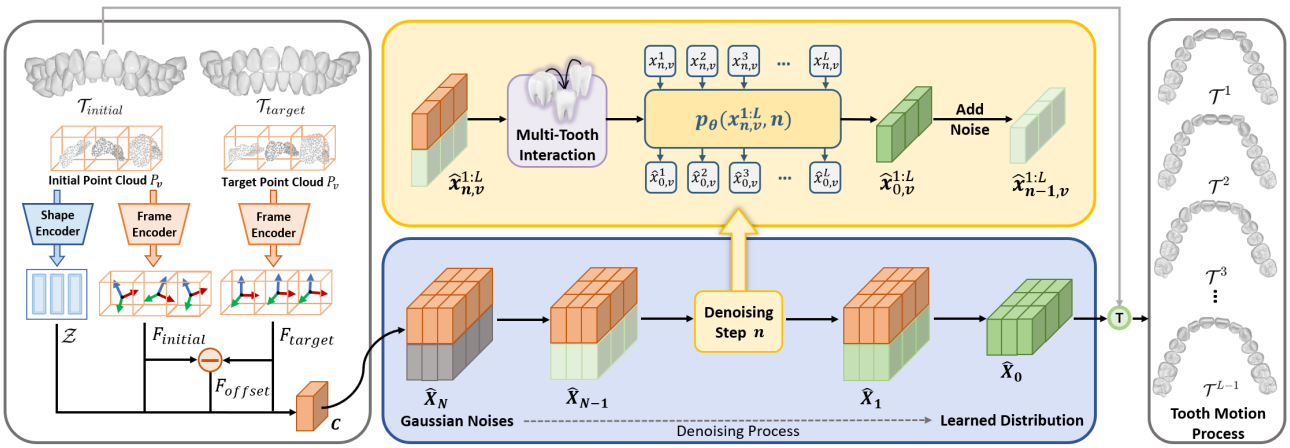


Figure 3: Overview of tooth motion generation via a conditional diffusion model. The input consists of point clouds representing initial $\mathcal{T}_{initial}$ and target $\mathcal{T}_{target}$ tooth alignments. The network initially obtains condition $c$ by encoding the tooth point cloud into frame and shape codes. Subsequently, we iteratively perform the denoising process. During each denoising step, we predict the original distribution using the trained denoiser $p_\theta$. After $N$ denoising steps, the motion distribution $\hat{\boldsymbol{X}}_0$ is acquired. Finally, we apply the learned $\hat{\boldsymbol{X}}_0$ to $\mathcal{T}_{initial}$ to obtain the intermediate tooth motion process.

captures the relative distance to the target frame. The condition $c_v$ is obtained by concatenating three encoded tooth frame features with shape latent code for the $v$-th tooth $c_v = \{F_{initial}, F_{offset}, F_{target}, \mathcal{Z}\}$.

Different from the previous UNet-based architecture on the 2D image, we build a simple GRU-based (Cho et al. 2014) denoising model on the tooth motion space, which is more suitable for the simplicity of the tooth motion features and more effective in capturing the sequential dependence of tooth motion with fewer parameters. The forward diffusion process gradually adds Gaussian noises to the original data $\boldsymbol{x}^{1:L}_{0,v} \sim q(\boldsymbol{x}^{1:L}_{0,v})$, as shown in Eq.(1):

$$q(\boldsymbol{x}^{1:L}_{n,v}|\boldsymbol{x}^{1:L}_{n-1,v}) := \mathcal{N}(\boldsymbol{x}^{1:L}_{n,v}; \sqrt{\alpha_n}\boldsymbol{x}^{1:L}_{n-1,v}, (1-\alpha_n)\mathbf{I}) \quad (1)$$

where $n$ is the diffusion step, the $\alpha_n \in (0,1)$ is a hyperparameter. There is no training in the noise-adding process.

To generate a whole intermediate motion process under $c_v$, we need to reverse the diffusion process. The denoising process can be approximated as a Markov chain with a learned mean and fixed variance in each step:

$$p_\theta(\boldsymbol{x}^{1:L}_{n-1,v}|\boldsymbol{x}^{1:L}_{n,v}, c_v) := \mathcal{N}(\boldsymbol{x}^{1:L}_{n-1,v}; \mu_\theta(\boldsymbol{x}^{1:L}_{n,v}, n, c_v), \sigma_n^2\mathbf{I}) \quad (2)$$

where $\theta$ represents the parameters of a neural network, and $c_v$ is a known condition. Learning the mean $\mu_\theta(\boldsymbol{x}^{1:L}_{n,v}, n, c_v)$ can be reparameterized as learning to predict the original data $\boldsymbol{x}^{1:L}_{0,v}$:

$$\mu_\theta = \frac{\sqrt{\alpha_n}(1-\bar{\alpha}_{n-1})\boldsymbol{x}^{1:L}_{n,v} + \sqrt{\bar{\alpha}_{n-1}}(1-\alpha_n)\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}^{1:L}_{n,v}, n, c_v)}{1-\bar{\alpha}_n} \quad (3)$$

where $\overline{\alpha}_t = \prod_{i=1}^{n} \alpha_i$. $\hat{\boldsymbol{x}}_\theta$ is the denoiser to predict $\boldsymbol{x}^{1:L}_{0,v}$. The training loss is defined as a reconstruction loss of $\boldsymbol{x}^{1:L}_{0,v}$:

$$\mathcal{L}_{\text{simple}} = E_{\boldsymbol{x}_{(0,v)},n}\|\boldsymbol{x}^{1:L}_{0,v} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{x}^{1:L}_{n,v}, n, c_v)\|_2^2 \quad (4)$$

We choose the initial input $\boldsymbol{x}^{1:L}_{0,v}$ as our learning objective (Song, Meng, and Ermon 2020) for two main reasons: 1) Incorporating constraints for better tooth motion generation. 2) Providing a strong anchor for denoising learning, ensuring accurate predictions, and avoiding deviations from the correct direction.

**Multi-tooth Interaction.** The multi-tooth interaction module aims to learn the relationships among teeth for each tooth at each time step. The motion of each tooth at each step is not independent, as they can be viewed as part of an interconnected multi-rigid-body system $\boldsymbol{X}_0 = \{\boldsymbol{x}^{1:L}_{0,v} | v \in \mathcal{V}\}$. It requires considering interactions among teeth, including adjacency, symmetry, and occlusal relationships. Forecasting the tooth motion process while comprehensively accounting for these factors constitutes a notably intricate endeavor. The graph neural networks are well-suited for this task. Inspired by TANet (Wei et al. 2020), we construct a multi-tooth interaction graph:

$$\mathcal{G} = (\boldsymbol{X}^l, \boldsymbol{E}) \quad (5)$$

where $\boldsymbol{X}^l$ represents tooth node features on motion step $l$. $\boldsymbol{E}$ contains three types of edges, representing adjacency, symmetry, and occlusal relationships of teeth, respectively. Meanwhile, we employ the Gated Graph Neural Network (Li et al. 2015) to abstract and aggregate the interaction information between teeth during multi-tooth motion from different feature levels at tooth alignment (local, relative global, and global) via three types of edges.

## Loss Function

In this section, we design multiple loss functions with two major components. One part is specifically designed for the tooth encoding module. The other part is the complementary soft constraints for the motion distribution.

**Tooth Model Encoding Loss** The tooth model encoding loss includes tooth frame encoding and tooth shape encoding.

$$\mathcal{L}_{encoding} = \lambda_1\mathcal{L}_{fe} + \lambda_2\mathcal{L}_{cd} \quad (6)$$

where $\lambda_1$ and $\lambda_2$ are balance factors.

**Tooth Frame Encoding Loss.** Given the point cloud of the tooth model $P_v$, we utilize the cosine similarity as the loss function to calculate the quaternion regression error in a point-wise prediction way as follows:

$$\mathcal{L}_{fe} = \frac{1}{|\mathcal{V}|M} \sum_{v \in \mathcal{V}} \sum_{j=0}^{M} (1 - cos(\boldsymbol{q}_{v,j}, \hat{\boldsymbol{q}}_{v,j})) \tag{7}$$

where $\hat{\boldsymbol{q}}_{v,j}$ is the predicted quaternion vector of the point $p_v \in P_v$ and $\boldsymbol{q}_{v,j}$ is the ground truth.

**Tooth Shape Encoding Loss.** To better capture the tooth shape latent code, we utilize the chamfer distance (CD) loss (Yuan et al. 2018) in the tooth shape encoder module. The loss function measures the difference between the decoder output $\hat{P}_v$ and the ground truth $P_v$. This loss is computed as:

$$\mathcal{L}_{cd} = \sum_{v \in \mathcal{V}} \Big( \sum_{\boldsymbol{p} \in P_v} \min_{\hat{\boldsymbol{p}} \in \hat{P}_v} ||\boldsymbol{p} - \hat{\boldsymbol{p}}||_2^2 + \sum_{\hat{\boldsymbol{p}} \in \hat{P}_v} \min_{\boldsymbol{p} \in P_v} ||\hat{\boldsymbol{p}} - \boldsymbol{p}||_2^2 \Big) \tag{8}$$

**Multi-tooth Motion Collaborative Constraints** In addition to $\mathcal{L}_{simple}$, we introduce multi-tooth collaborative constraints at two levels: frame-level and point-cloud-level. These constraints apply the predicted motion $(\hat{\mathbf{r}}, \hat{\mathbf{t}})$ learned by the denoiser to tooth frames and point clouds, with direct supervision on both. Overall, our motion loss is:

$$\mathcal{L}_{motion} = \lambda_3 \mathcal{L}_{simple} + \lambda_4 \mathcal{L}_{frame} + \lambda_5 \mathcal{L}_{pc} + \lambda_6 \mathcal{L}_{col} \tag{9}$$

where $\lambda_3$ to $\lambda_6$ are balance factors. $\mathcal{L}_{simple}$ is the reconstruction loss, and $\mathcal{L}_{frame}$, $\mathcal{L}_{pc}$, and $\mathcal{L}_{collision}$ are the constraint losses.

**Frame Constraint Loss.** $\mathcal{L}_{frame}$ constrains the posture and position of the tooth model to help the network learn tooth motion in spatial space. This is defined as:

$$\mathcal{L}_{frame} = \frac{1}{|\mathcal{V}|(L-1)} \sum_{v \in \mathcal{V}} \sum_{l=1}^{L-1} (\left\| \hat{\boldsymbol{q}}_v^l - \boldsymbol{q}_v^l \right\|_2 + \left\| \hat{\boldsymbol{o}}_v^l - \boldsymbol{o}_v^l \right\|_2) \tag{10}$$

where $\boldsymbol{q}_v^l$ is the ground truth quaternion representation of three axes $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ in $F_v$ at the motion step $l$. The $\hat{\boldsymbol{o}}_v^l = \boldsymbol{o}_v^{l-1} + \hat{\mathbf{t}}_v^l$ and the $\hat{\boldsymbol{q}}_v^l = \hat{R}_v^l \hat{\boldsymbol{q}}_v^{l-1}$, where $\hat{R}_v^l = e^{\hat{\mathbf{r}}_v^l}$ by Rodrigues' formula (Gray 1980). Since the $F_{initial}$ is known, $\hat{\boldsymbol{q}}^{l-1}$ and $\boldsymbol{o}^{l-1}$ can be obtained.

**Tooth and Jaw Constraint Loss.** The tooth point cloud information can provide finer data to constrain the occlusal relationships between teeth. We introduce the chamfer vector loss between point cloud sets using smooth L1 loss:

$$\mathcal{L}_{pc} = \sum_{(a,b) \in \kappa} \sum_{l=1}^{L-1} \| \boldsymbol{V}_{\hat{P}_a, \hat{P}_b}^l - \boldsymbol{V}_{P_a, P_b}^l \|_{\mathcal{S}} \tag{11}$$

where $\boldsymbol{V}_{P_a,P_b}^l$ is the chamfer vector between two ground truth point sets $P_a$ and $P_b$ in step $l$, which defined follow (Wei et al. 2020). The $\kappa$ represents the set pairs, including the pairs of adjacent teeth and all points in the upper and lower jaw. $\hat{P}^l = \hat{R}^l(\hat{P}^{l-1} - \boldsymbol{o}^{l-1}) + \boldsymbol{o}^{l-1} + \hat{\boldsymbol{t}}^l$ represents the predicted tooth point cloud.

**Collision Avoiding Loss.** The Lennard-Jones potential (Hansen and Verlet 1969) is a model that describes the interactions between atoms with the property of close repulsion and distant alienation. Inspired by it, $\mathcal{L}_{col}$ is introduced to control the distance between the adjacent tooth to make teeth collision-free and as close as possible. It is defined as:

$$\mathcal{L}_{col} = \sum_{(a,b) \in \kappa} \sum_{l=1}^{L-1} \left( \left( \frac{1}{1 + d^l / s^l} \right)^{12} - 2 \left( \frac{1}{1 + d^l / s^l} \right)^{6} \right) \tag{12}$$

where $d^l = d_{np}^l + \delta$, $d_{np}^l$ represents the nearest point pair distance between predicted $\hat{P}_a^l$ and $\hat{P}_b^l$. To avoid errors due to $\hat{P}_a^l$ and $\hat{P}_b^l$ overlap, we add $\delta$ to constrain $d_{np}^l$ ensuring a minimum non-overlapping distance between them. $\delta$ is the empirical parameter. $s$ is the distance between their centroids.

# Experiments

In this section, we show quantitative comparisons and illustrate our qualitative results. We also conduct ablation studies to analyze the performance of our method, as well as the design choices in our model.

## Dataset and Implementation Details

Our dataset consists of 1050 dental cases. Each dental case includes a group of models before and after alignment, the corresponding whole tooth motion process, and the tooth frames. These data are collected from hospitals and labeled by professional dentists. The labeling of the tooth motion process includes generating intermediate tooth frames through interpolation based on initial and target tooth frames using an auxiliary labeling system. The dentist then manually adjusts tooth position and posture at each step. The max motion generation step is $L = 20$. We randomly divided the data into three parts for network training: 787 for training, 105 for validation, and 158 for testing. Our model is trained with $N = 100$ noising steps and a cosine noise schedule. All of them are trained on a single NVIDIA RTX 4090 GPU. We use zero padding to deal with the missing teeth and downsample $M = 400$ points for each tooth point cloud from the intraoral scan model by farthest point sampling. The dimension of the tooth latent shape code $\mathcal{Z}$ is $z = 16$. The denoising network is a GRU with four hidden layers and a latent dimension 256. The empirical parameter $\delta$ is 0.1. We weigh our loss terms by $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10$, $\lambda_4 = 1$, $\lambda_5 = 0.1$, and $\lambda_6 = 0.1$. The framework is trained in two stages. The encoding part was trained first, and then its output was used as input for training the motion generation part.

## Evaluation Metrics

To evaluate the performance of our method, we modify three evaluation metrics to make it suitable for tooth motion process prediction. We use $R_{error}$ and $T_{error}$ (Sattler et al. 2018) to evaluate the average error of rotation and translation of each tooth in all motion steps, respectively. These are calculated by:

$$R_{error} = \frac{1}{|\mathcal{V}|} \frac{1}{L} \sum_{v \in \mathcal{V}} \sum_{l=1}^{L} (\mathrm{Tr}((R_v^l)^{-1} \hat{R}_v^l) - 1) \tag{13}$$

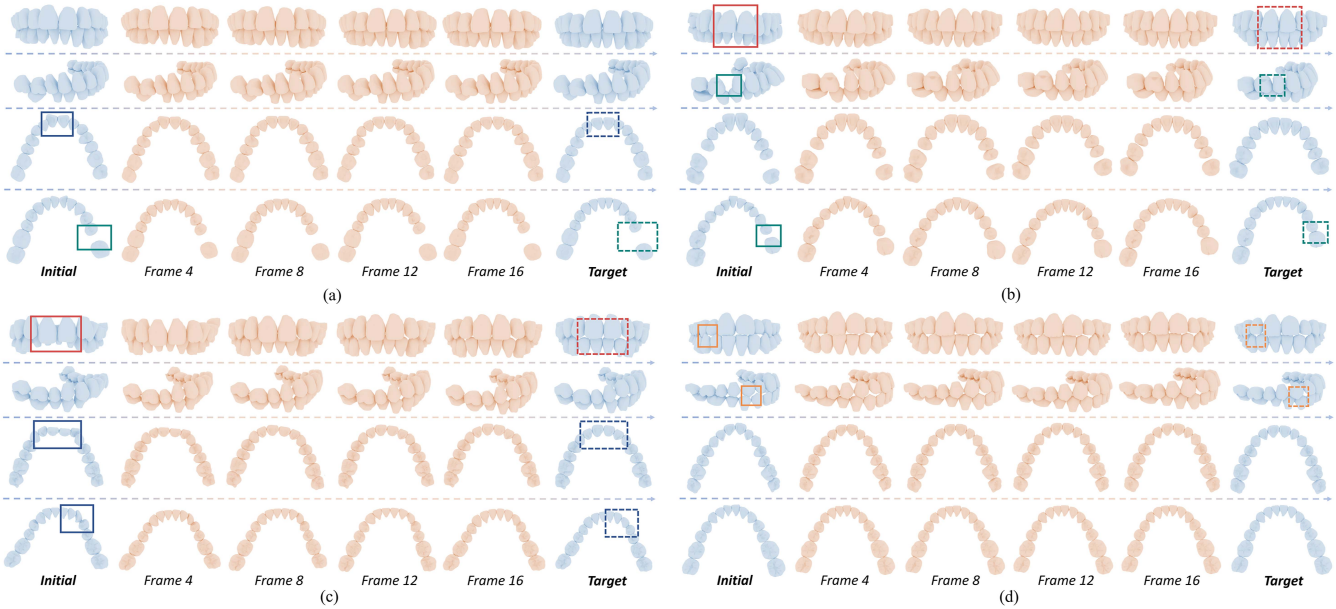


Figure 4: Visualization of the results of our network. (a)-(d) Four cases from our test dataset. In each case, six columns represent the input initial alignment (blue), the output motion process of our network (orange), and the target alignment (blue). The four rows from top to bottom are the front view, side view, upper and lower view.

$$T_{error} = \frac{1}{|\mathcal{V}|}\frac{1}{L}\sum_{v\in\mathcal{V}}\sum_{l=1}^{L}\left\|\hat{\mathbf{t}}_v^l - \mathbf{t}_v^l\right\|_2 \tag{14}$$

$\text{AUC}_{\text{point}}$ (Wei et al. 2020) represents the point reconstruction error of each tooth per step. It means the area under the PCT curve of ADD which calculates the mean point-wise distance between the predicted point cloud and ground truth (Hinterstoisser et al. 2013).

## Qualitative Results

Fig. 4 shows our method can generate excellent motion processes under different malocclusion problems. The blue, green, red, and orange rectangular boxes (solid) focus on the misalignment of initial teeth caused by dental crowding, missing teeth, overbite, and underbite problems, respectively. The rectangular boxes (dotted) focus on the abnormal changes in the target alignment after the whole intermediate motion process. We can also handle the middle line that is not aligned in Fig. 4a, the overjet in Fig. 4b, and the dental arch narrowing in Fig. 4c. We refer to the website for related demo results: https://yeyingstudy.github.io/.

**Variable-length motion.** Our model enables concurrently generating the motion process of all timesteps. The step of motion generation is determined by the size of the input $\boldsymbol{x}^{1:L} \sim \mathcal{N}(0, \mathbf{I})$ that initializes the denoising process, allowing for the variable-length motion process (Janner et al. 2022). $\boldsymbol{x}^{1:L}$ can be viewed as the seed process.

**Inference time.** We adopted Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2020) and used the proposed tooth representation to accelerate sampling. For a tooth alignment taking 20 motion steps for example, the inference time requires 10.25 seconds on average.

| | $R_{error}\downarrow$ | | $T_{error}\downarrow$ | | $\text{AUC}_{\text{point}}\uparrow$ | |
|---|---|---|---|---|---|---|
| Lengths | 5 | 20 | 5 | 20 | 5 | 20 |
| Interp | 0.36 | 0.29 | 0.128 | 0.074 | 81.95 | 84.95 |
| TANet | 3.24 | - | 3.709 | - | 75.89 | - |
| Regression | 0.58 | 0.44 | 0.072 | 0.043 | 84.63 | 86.53 |
| MDM | 0.78 | 0.46 | 0.092 | 0.067 | 83.23 | 75.39 |
| Our | **0.36** | **0.18** | **0.069** | **0.039** | **90.05** | **93.06** |

Table 1: Results comparison of different methods. The coordinate unit is *mm* for $T_{error}$, *degree* for $R_{error}$.

## Comparisons

To the best of our knowledge, we are the first to generate tooth motions in a learning-based manner. To verify the effectiveness of our method, we chose and designed four comparative experiments. **Interpolation** (Chapuis et al. 2022): uses linear and quaternion spherical interpolation for tooth position and posture. **TANet** (Wei et al. 2020): predicts the tooth motion step by step by a shared-parameter point cloud network. **Regression**: we use the initial tooth frame concatenated with the target frame as input for our denoiser network to regress tooth motion directly. **MDM** (Tevet et al. 2022): we apply the same image-inpainting strategy for our task.

Tab. 1 shows our network could achieve the highest accuracy in three metrics. The interpolation method can only create smooth transitions without many motion details. The experimental results indicate that TANet allows for predicting up to 5 steps. In contrast, we can predict 20 steps based on our proposed tooth representation and network method. The MDM strategy is not as good as our performance re-

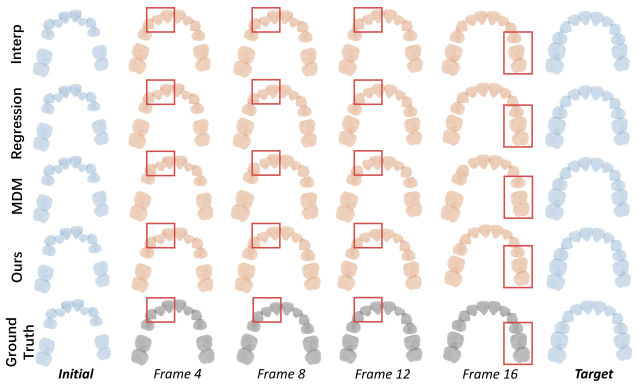


Figure 5: Qualitative comparison with the other methods. The blue represents the known alignment, the orange represents the output motion process of these methods, and the gray represents the ground truth. The red boxes denote the compare areas.

sults and has a serious collision due to the lack of relevant constraint restrictions. It's worth noting that the regression method validates the tooth frames' capability to represent tooth position and posture information. Fig. 5 shows that our method generates a more accurate tooth motion process.

## Ablation Study

To validate the effects of our network components, we conducted ablation experiments by augmenting the base network. The base network represents the network trained only with $\mathcal{L}_{simple}$. The results are shown in Tab.2.

**Effect of Tooth Latent Representation.** We explore the effect of two key components in tooth latent representation. 1) To verify the effectiveness of the tooth frame in representing the tooth posture and position, we remove the tooth frame from the base network. The accuracy is greatly reduced compared to the base network (the first row). As the network learned only the intermediate motion process without knowledge of the transformation condition, removing the tooth frame from the base network results in its degradation into unconditional diffusion. 2) The result shows that the tooth shape latent code can improve the accuracy of the base network (the third row). The shape code succinctly captures the tooth point cloud morphology, reducing computations and introducing inter-tooth constraint adjustments in multi-tooth motion for aiding network learning.

**Effect of Multi-tooth Collaboration.** We augment the base network with the interaction module and loss constraints to validate the effect of the multi-tooth collaboration. The results in Tab.2 indicate that each constraint contributes to enhancing our base network. 1) That is because, through the multi-tooth interaction module, our network can better capture adjacency, symmetry, and occlusal information among teeth at the feature level. 2) The multi-tooth loss function can act as a kind of soft constraint to guide the generated motion distribution to adhere to the correct posture and position of $\mathcal{L}_{frame}$, the occlusal relationship of $\mathcal{L}_{pc}$, and the distance control between teeth of $\mathcal{L}_{col}$.

| Method | $R_{error} \downarrow$ | $T_{error} \downarrow$ | $\text{AUC}_{\text{point}} \uparrow$ |
|---|---|---|---|
| w/o Frame | 0.456 | 0.0941 | 65.1907 |
| Base | 0.198 | 0.0497 | 90.9010 |
| Base + Shape Code | 0.181 | 0.0403 | 92.2750 |
| Base + Interaction | 0.184 | 0.0415 | 92.1020 |
| Base + $\mathcal{L}_{frame}$ | 0.186 | 0.0445 | 92.7169 |
| Base + $\mathcal{L}_{pc}$ | 0.191 | 0.0403 | 92.6158 |
| Base + $\mathcal{L}_{col}$ | 0.183 | 0.0408 | 92.0803 |
| Our | 0.178 | 0.0399 | 93.0670 |
| Our* | **0.178** | **0.0386** | **93.2190** |

Table 2: Ablation study of tooth latent representation and multi-tooth collaboration. Our* represents the result performance of post-processing. Taking $L = 20$ as an example.

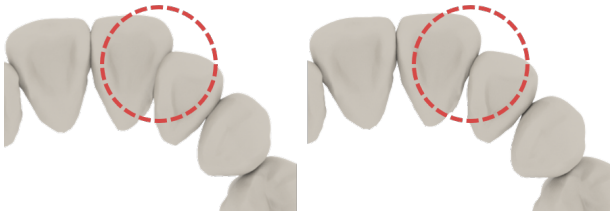

Figure 6: The circle focus on the post processing region. Left and right show the output of our network and the results after post processing.

## Post Processing

The neural network learning uses a soft constraint, which leads to the possibility that its outputs may still have a slight collision. In post-processing, we address teeth collisions by using detailed distance controls to ensure no collisions completely as shown in Fig. 6. Specifically, the goal is to minimize the energy function $E_{col}$. The $E_{col}$ has the same form as $\mathcal{L}_{col}$. The difference is that $s$ is the summarization of max SDF values of $\mathcal{T}_a$ and $\mathcal{T}_b$, and $d = d_{SDF}$ is the lowest signed distance value between the surface of $\mathcal{T}_a$ and $\mathcal{T}_b$. Compared to $d_{np}$, $d_{SDF}$ offers a more accurate distance with the signed, effectively avoiding the problem caused by the tooth point cloud overlap. Therefore, we use $d_{SDF}$ in the output result of our network to enhance our results further.

## Conclusion

In this paper, we present the first diffusion-based approach for tooth motion generation, treating it as a multi-tooth motion distribution fitting problem. Our tooth latent representation effectively captures tooth position, posture, and shape features and acts as conditional guidance. The proposed multi-tooth interaction module and multiple constraint losses establish the motion collaboration among teeth. Our work demonstrates that modeling a multi-tooth motion process as a motion distribution is an effective solution for tooth motion generation results.

## Acknowledgments

This project is partially supported by the National Key R&D Plan on Strategic International Scientific and Technological Innovation Cooperation Special Project (No.2021YFE0203800), the National Natural Science Foundation of China under Grant (No.62172257), the GRF 17210419 and GRF 17212120 by Research Grant Council of Hong Kong.

## References

Chapuis, M.; Lafourcade, M.; Puech, W.; Guillerm, G.; and Faraj, N. 2022. Animating and Adjusting 3D Orthodontic Treatment Objectives. In *GRAPP 2022-17th International Conference on Computer Graphics Theory and Applications*, 60–67. SCITEPRESS.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cui, Z.; Fang, Y.; Mei, L.; Zhang, B.; Yu, B.; Liu, J.; Jiang, C.; Sun, Y.; Ma, L.; Huang, J.; et al. 2022. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nature communications*, 13(1): 2096.

Cui, Z.; Li, C.; Chen, N.; Wei, G.; Chen, R.; Zhou, Y.; Shen, D.; and Wang, W. 2021. TSegNet: An efficient and accurate tooth segmentation network on 3D dental model. *Medical image analysis*, 69: 101949.

Dabral, R.; Mughal, M. H.; Golyanik, V.; and Theobalt, C. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9760–9770.

Gray, J. J. 1980. Olinde Rodrigues' paper of 1840 on transformation groups. *Archive for History of Exact Sciences*, 21(4): 375–385.

Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; and Lu, J. 2022. Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17113–17122.

Guo, W.; Bie, X.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2022. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13053–13064.

Hansen, J.-P.; and Verlet, L. 1969. Phase transitions of the Lennard-Jones system. *Physical Review*, 184(1): 151.

Harvey, F. G.; Yurick, M.; Nowrouzezahrai, D.; and Pal, C. 2020. Robust motion in-betweening. *ACM Transactions on Graphics*, 39(4): 60–1.

Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; and Navab, N. 2013. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision*, 548–562. Springer.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A. A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video Diffusion Models. In *NeurIPS*.

Janner, M.; Du, Y.; Tenenbaum, J.; and Levine, S. 2022. Planning with diffusion for flexible behavior synthesis. arXiv:2205.09991.

Jiang, C.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; Anguelov, D.; et al. 2023. MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9644–9653.

Kaufmann, M.; Aksan, E.; Song, J.; Pece, F.; Ziegler, R.; and Hilliges, O. 2020. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision*, 918–927. IEEE.

Kim, J.; Byun, T.; Shin, S.; Won, J.; and Choi, S. 2022. Conditional motion in-betweening. *Pattern Recognition*, 132: 108894.

Li, J.; Liu, K.; and Wu, J. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17142–17151.

Li, X.; Bi, L.; Kim, J.; Li, T.; Li, P.; Tian, Y.; Sheng, B.; and Feng, D. 2020a. Malocclusion treatment planning via pointnet based spatial transformation network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III*, 105–114. Springer.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Li, Z.; Li, K.; and Li, B. 2009. Research on path planning for tooth movement based on genetic algorithms. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, 421–424. IEEE.

Li, Z.; Liu, T.; Li, H.-A.; and Sun, Z. 2020b. Orthodontic path planning method based on optimized artificial bee colony algorithm. In *Journal of Physics: Conference Series*, volume 1544, 012017. IOP Publishing.

Li, Z.; and Yang, G. 2011. Research on Simulation and Optimization Method for Tooth Movement in Virtual Orthodontics. In *Advances in Computer Science, Environment, Ecoinformatics, and Education: International Conference, CSEE 2011, Wuhan, China, August 21-22, 2011. Proceedings, Part I*, 270–275. Springer.

Lyu, Z.; Kong, Z.; Xu, X.; Pan, L.; and Lin, D. 2022. A Conditional Point Diffusion-Refinement Paradigm for 3D Point Cloud Completion. In *The Tenth International Conference*

*on Learning Representations, 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ma, Q.; Wei, G.; Zhou, Y.; Pan, X.; and Wang, W. 2020. SR-Fet: Spatial Relationship Feature Network for Tooth Point Cloud Classification. *Computer Graphics Forum*, 39(7): 267–277.

Ma, T.; Lyu, J.; Yang, Q.; Li, Z.; Li, Y.; Chen, Y.; and Ren, X. 2021. Orthodontic Overcorrection Scheme Generation Based on Improved Multiparticle Swarm Optimization. *Journal of Healthcare Engineering*, 2021: 1–12.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning, 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qin, J.; Zheng, Y.; and Zhou, K. 2022. Motion In-betweening via Two-stage Transformers. *ACM Transactions on Graphics*, 41(6): 1–16.

Qiu, L.; Ye, C.; Chen, P.; Liu, Y.; Han, X.; and Cui, S. 2022. Darch: Dental arch prior-assisted 3d tooth instance segmentation with weak annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20752–20761.

Raab, S.; Leibovitch, I.; Tevet, G.; Arar, M.; Bermano, A. H.; and Cohen-Or, D. 2023. Single Motion Diffusion.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. 2018. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 8601–8610.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2023. Human Motion Diffusion as a Generative Prior.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models.

Song, W.; Liang, Y.; Yang, J.; Wang, K.; and He, L. 2021. Oral-3d: Reconstructing the 3d structure of oral cavity from panoramic x-ray. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 566–573.

Tang, X.; Wang, H.; Hu, B.; Gong, X.; Yi, R.; Kou, Q.; and Jin, X. 2022. Real-time controllable motion transition for characters. *ACM Transactions on Graphics*, 41(4): 1–10.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2022. Human Motion Diffusion Model. arXiv:2209.14916.

Wang, C.; Wei, G.; Wei, G.; Wang, W.; and Zhou, Y. 2022. Tooth Alignment Network Based on Landmark Constraints and Hierarchical Graph Structure. *IEEE Transactions on Visualization and Computer Graphics*, 1–12.

Wang, H.; Wu, Y.; Guo, S.; and Wang, L. 2023. PDPP:Projected Diffusion for Procedure Planning in Instructional Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14836–14845.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5): 1–12.

Wei, G.; Cui, Z.; Liu, Y.; Chen, N.; Chen, R.; Li, G.; and Wang, W. 2020. TANet: towards fully automatic tooth arrangement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, 481–497. Springer.

Wei, G.; Cui, Z.; Zhu, J.; Yang, L.; Zhou, Y.; Singh, P.; Gu, M.; and Wang, W. 2022. Dense representative tooth landmark/axis detection network on 3D model. *Computer Aided Geometric Design*, 94: 102077.

Wu, W.; Qi, Z.; and Fuxin, L. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 9621–9630.

Wyatt, J.; Leach, A.; Schmon, S. M.; and Willcocks, C. G. 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 650–656.

Xu, S.; Wang, Y.-X.; and Gui, L. 2023. Stochastic Multi-Person 3D Motion Forecasting. In *The Eleventh International Conference on Learning Representations*.

Yang, L.; Shi, Z.; Wu, Y.; Li, X.; Zhou, K.; Fu, H.; and Zheng, Y. 2020. iOrthoPredictor: model-guided deep prediction of teeth alignment. *ACM Transactions on Graphics*, 39(6): 216.

Yf, A.; Qian, M. A.; Gw, A.; Zc, B.; A, Y. Z.; and Wwb, C. 2022. TAD-Net: tooth axis detection network based on rotation transformation encoding. *Graphical Models*, 121: 101138.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. Pcn: Point completion network. In *2018 international conference on 3D vision*, 728–737. IEEE.

Zhang, C.; Elgharib, M.; Fox, G.; Gu, M.; Theobalt, C.; and Wang, W. 2022. An Implicit Parametric Morphable Dental Model. *ACM Transactions on Graphics*, 41(6): 1–13.