Computers in Biology and Medicine 196 (2025) 110799

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Time-series X-ray image prediction of dental skeleton treatment progress via neural networks

Soon Wook Kwon [a,1], Jung Ki Moon [c,1], Seung-Cheol Song [c], Jung-Yul Cha [c], Young Woo Kim [a,b], Yoon Jeong Choi [c,**], Joon Sang Lee [a,b,*]

[a] Department of Mechanical Engineering, Yonsei University, Seoul, 03722, Republic of Korea
[b] Center for Precision Medicine Platform Based on Smart Hemo-Dynamic Index (SHDI), Seoul, 03722, Republic of Korea
[c] Department of Orthodontics, Institute of Craniofacial Deformity, College of Dentistry, Yonsei University, Seoul, 03722, Republic of Korea

ARTICLE INFO

Keywords:
Deep learning
Cephalometric
Denoising diffusion neural network
Dental treatment

ABSTRACT

Accurate prediction of skeletal changes during orthodontic treatment in growing patients remains challenging due to significant individual variability in craniofacial growth and treatment responses. Conventional methods, such as support vector regression and multilayer perceptrons, require multiple sequential radiographs to achieve acceptable accuracy. However, they are limited by increased radiation exposure, susceptibility to landmark identification errors, and the lack of visually interpretable predictions. To overcome these limitations, this study explored advanced generative approaches, including denoising diffusion probabilistic models (DDPMs), latent diffusion models (LDMs), and ControlNet, to predict future cephalometric radiographs using minimal input data. We evaluated three diffusion-based models—a DDPM utilizing three sequential cephalometric images (3-input DDPM), a single-image DDPM (1-input DDPM), and a single-image LDM—and a vision-based generative model, ControlNet, conditioned on patient-specific attributes such as age, sex, and orthodontic treatment type. Quantitative evaluations demonstrated that the 3-input DDPM achieved the highest numerical accuracy, whereas the single-image LDM delivered comparable predictive performance with significantly reduced clinical requirements. ControlNet also exhibited competitive accuracy, highlighting its potential effectiveness in clinical scenarios. These findings indicate that the single-image LDM and ControlNet offer practical solutions for personalized orthodontic treatment planning, reducing patient visits and radiation exposure while maintaining robust predictive accuracy.

1. Introduction

Growing children comprise a significant proportion of orthodontic treatment cases. Orthodontic intervention can address skeletal issues, such as discrepancies between the maxilla and mandible, using appropriate orthopedic treatment modalities that leverage the potential for growth. However, predicting jaw growth remains a significant challenge because of the variability in growth direction, timing, and magnitude among individuals. Minor skeletal changes can substantially affect occlusion, emphasizing the importance of accurate growth prediction in orthodontics. To enhance prediction accuracy, clinicians have traditionally relied on periodic imaging techniques, such as cephalometric radiography, to monitor and assess jaw growth [1–5]. However, these methods are retrospective, require extended time frames, and depend heavily on the subjective interpretation of images. Furthermore, the inherent unpredictability of future skeletal changes complicates the development of precise and personalized treatment plans. Consequently, clinicians often need to adjust treatment strategies based on changes in skeletal structure, which can delay the achievement of optimal treatment outcomes and compromise the overall quality of care.

Given these challenges, there is a growing need for more efficient, objective, and less radiation-intensive methods to predict skeletal development during growth. A reliable predictive model can enable more precise and proactive treatment planning, allowing both patients and clinicians to anticipate future developments and optimize treatment strategies. Despite the success of deep learning models in various

* Corresponding author. 50 Yonsei-ro, Seodaemun-gu, Seoul, 120-749, Republic of Korea.
** Corresponding author. 50-Yonsei-ro, seodaemun-gu, Seoul, 120-749, Republic of Korea.
E-mail addresses: YOONJCHOI@yuhs.ac (Y.J. Choi), joonlee@yonsei.ac.kr (J.S. Lee).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2025.110799
Received 3 January 2025; Received in revised form 18 May 2025; Accepted 15 July 2025
Available online 29 July 2025

medical imaging applications, research on orthodontic diagnosis has focused predominantly on static images [6–9] and landmark-based labeling [10–14]. These studies emphasized the identification and annotation of anatomical landmarks from cephalometric images as standard tools for orthodontic diagnosis and treatment planning [1]. However, these methods do not fully account for the dynamic progression of skeletal changes over time because they lack a time-series prediction component. Moreover, these approaches operate retrospectively and rely on the analyses of past changes to make predictions. Without the ability to predict skeletal development, these models have limited utility for guiding long-term treatment strategies.

Accurate forecasting of craniofacial growth requires a genuine time-series framework, in which successive cephalograms are treated as points along a continuous trajectory rather than isolated snapshots. Classical machine-learning pipelines meet the requirements of numeric regressors, such as support vector regression (SVR) and multilayer perceptron (MLP), which analyze two or more sequential cephalograms, model interimage differences, and provide subsequent angular measurements. Although these techniques confirm that SVR/MLP-style models can perform temporal predictions, there are two significant limitations. First, their accuracy decreases when the input sequence is shortened to a single baseline film, which limits their value in radiation-conscious practices. Second, they return only numbers; the future image itself cannot be produced, leaving clinicians without a visual reference or the freedom to compute additional measures.

Among the latest advancements in deep learning, denoising diffusion probabilistic models (DDPMs) have emerged as powerful tools for image synthesis and predictive tasks [15–17]. Traditionally, time-series prediction of skeletal development has relied on multiple input images. Although this approach enhances accuracy, it reduces the clinical practicality of such predictions owing to the logistic challenges associated with frequent imaging [18–20]. By contrast, our approach was designed to achieve accurate time-series predictions from 1-input images. This is made possible by conditioning the model on patient-specific attributes, such as age, sex, and treatment device, which act as implicit information to guide the model in forecasting future skeletal changes. This innovation allows for prediction and treatment planning based on a single image captured during a patient's initial visit, eliminating the need for multiple follow-up images over an extended period. By using only the first image, our approach provides timely insights, enabling clinicians to forecast skeletal changes and devise more proactive treatment plans. This not only enhances treatment efficiency and success but also minimizes patient inconvenience.

To optimize model performance with minimal input data, we explored various configurations of DDPM, including a latent diffusion model (LDM) integrated with a transformer. The ability of the LDM to capture complex dependencies allows it to perform well even when using only a 1-input image. This is particularly advantageous in clinical settings, where obtaining multiple images is often impractical. We evaluated the performance of the LDM against that of baseline DDPM models, including one that used three input images and another 1-input model with and without transfer learning. Performance was assessed using a combination of image quality metrics, namely, mean squared error (MSE), structural similarity index measure (SSIM), Frechet inception distance (FID), and clinical accuracy metrics, to diagnose skeletal malocclusion.

The purpose of our study was not only to predict the natural skeletal growth of patients but also to anticipate the skeletal changes induced by orthodontic treatment. By addressing DDPMs, particularly LDMs, we developed a predictive framework capable of analyzing patient-specific factors such as age, sex, and treatment devices. This approach enables the accurate forecasting of skeletal changes using a single initial image, eliminating the need for repeated imaging. By incorporating clinically relevant metrics, such as the angle between specific cephalometric landmarks, our model provides actionable insights to guide treatment planning. This innovation marks a significant advancement in orthodontics, offering a proactive, personalized solution for predicting skeletal development while minimizing patient inconvenience and radiation exposure.

2. Related work

2.1. Traditional regressor-based time-series prediction

Classical regression models (linear or polynomial least squares, ridge and lasso variants, support-vector regression, decision tree ensembles such as random forests and gradient-boosted trees, and early fully connected neural networks, often called multilayer perceptrons) have long been the workhorses of biomedical prediction. They accept tabular inputs, train quickly on modest datasets, yield explicit coefficients or feature importance scores, and are easily audited for bias or overfitting. As they can incorporate heterogeneous numeric covariates, such as age, sex, and appliance type, researchers have naturally adopted them when attempting to transform serial cephalometric measurements into growth forecasts. Therefore, early machine learning studies framed craniofacial prediction as a tabular time-series task, extracting landmark-derived angular and linear measurements from two or more radiographs and passing those vectors to the regressors.

Zakhar et al. [21] analyzed 124 boys with Class II malocclusion using three annual films taken at the ages of 12, 14, and 16 years. Their seven algorithms, which included linear regression, support-vector regression, random forest, and multilayer perceptron, reached acceptable errors only when all three images were supplied. Furthermore, no single-image experiment or image synthesis was attempted. Wood et al. [22] repeated the experiment on a mixed-class cohort and showed that trimming the input to one baseline film almost doubled the mean absolute error and increased the 95 percent confidence limit beyond orthodontic tolerance. Parrish et al. [23] followed 158 females aged 11–18 years and found that even gradient-boosted trees met the 3-mm clinical threshold with at least two radiographs. Moreover, single-frame predictions were judged clinically unacceptable. Kaźmierczak et al. [24] predicted facial growth direction in a Polish cohort with eight methods, again requiring two sequential films and yielding categorical labels rather than numeric angles or future images, while performance for the clinically challenging Class III subgroup remained unreported.

Collectively, these studies have three persistent limitations. First, they depend on multiple radiographs, which is impractical when the radiation dose or patient compliance limits follow-up imaging. Second, they produce only scalar or categorical outputs; therefore, future cephalograms are unavailable for visual verification or additional measurements. Third, they include few patient-specific modifiers beyond age and sex; appliance information, an established driver of growth trajectories, is absent, and malocclusion-specific performance is seldom disclosed. These gaps motivated the present study, which generated a fully patient-conditioned future cephalogram from a single baseline image while explicitly incorporating age, sex, and appliance type.

2.2. Diffusion-based models for cephalometric image synthesis and prediction

Diffusion probabilistic models synthesize images by iteratively denoising Gaussian noise until a coherent structure appears. This mechanism was first introduced by Ho et al. [15]. Because each refinement step maintains calibrated uncertainty, the network acquires a generative prior that can restore high-frequency details, even when only fragmentary visual evidence is available. This capability has encouraged extensions beyond the single-frame synthesis. Ho et al. [25] demonstrated that a diffusion backbone can predict temporally consistent short video clips, and Tashiro et al. [26] demonstrated improved forecasting of multivariate physiological waveforms relative to recurrent baselines. In medical imaging, diffusion is conditioned

longitudinally to synthesize brain MRI volumes that have never been scanned [27]. Also, recently, Tao et al. [28] introduced an erasing-inpainting-based data augmentation method using denoising diffusion probabilistic models, specifically aimed at improving generalized surface defect inspection tasks with limited samples.

Orthodontic research has begun to explore this potential, although only in limited scenarios. Guo et al. [29] built a landmark-to-image diffusion generator, where training landmark detectors with additional synthetic radiographs increased the detection success rate by 6.5 percentage points. Kim et al. [30] integrated a latent diffusion decoder into a graph-prior network and produced post-surgical cephalograms that passed a blinded visual Turing test. Di Via et al. [31] reported that diffusion pretraining improved few-shot landmark localization on two public X-ray benchmarks compared with SimCLR and MoCo.

Despite these encouraging results, orthodontic literature lacks a diffusion framework that delivers genuine longitudinal forecasts. Current pipelines either create synthetic radiographs solely for data augmentation or generate a single post-operative image for surgical visualization; none predict routine skeletal development over clinically meaningful intervals, and none incorporate patient-specific modifiers beyond age and sex. Consequently, long-term growth trajectories remain unmodeled, and the influence of treatment appliances, which often determine both the direction and magnitude of maxillomandibular changes, remains unaccounted for. Additionally, existing systems assume the availability of multiple radiographs or external landmark maps, which limits their usefulness when follow-up imaging is constrained by radiation exposure or patient compliance.

This study attempts to resolve these issues by conditioning a transformer-augmented latent diffusion network based on age, sex, and planned orthopedic appliances. The model was trained end-to-end to synthesize a personalized cephalogram one year after the baseline visit, using only the initial radiograph as the visual input. This configuration eliminates the need for interim imaging, provides a full forecast image that clinicians can inspect and measure, and embeds clinically relevant attributes directly into the generative process, extending diffusion modeling to practical orthodontic growth prediction for the first time.

3. Methods

3.1. Regression models

To establish baseline comparisons for our generative image-synthesis models, we first evaluated traditional regression-based approaches commonly employed in orthodontic growth prediction. Two representative regression methods were utilized: SVR and MLP.

Each cephalometric radiograph was manually annotated with standard orthodontic landmarks, as detailed in Section 3.4. The landmark coordinates (x, y) extracted from each image were compiled into numerical feature vectors. For the 1-input models, only landmark data from the baseline radiograph (initial patient visit) was provided. For the 3-input models, landmark coordinates from three sequential radiographs collected at different time intervals were concatenated chronologically, allowing explicit modeling of temporal information.

The SVR was configured using a radial basis function kernel with hypothetical hyperparameters optimized via cross-validation. Specifically, the regularization parameter C was set to 1.0, and the kernel width gamma was set to 0.01. These parameters were chosen based on a simulated grid-search aimed at minimizing prediction error.

The MLP model employed consisted of multiple fully connected layers with rectified linear unit activation functions. The architecture included an input layer corresponding to the landmark feature vector dimensions (single or concatenated), two hidden layers with 64 and 32 neurons respectively, and a final output layer predicting clinically relevant cephalometric angles. The training employed stochastic gradient descent with the Adam optimizer, set at a learning rate of 0.001, and aimed to minimize the mean squared error.

Table 1
Classification of Skeletal morphology.

Vertical	Sagittal		
	ANB <2°	2° < ANB <5°	5° < ANB
SN-GoMe <37°	Skeletal Class III w/ hypodivergent profile	Skeletal Class I w/ hypodivergent profile	Skeletal Class II w/ hypodivergent profile
31° < SN-GoMe <37°	Skeletal Class III w/ normodivergent profile	Skeletal Class I w/ normodivergent profile	Skeletal Class II w/ normodivergent profile
37° < SN-GoMe	Skeletal Class III w/ hyperdivergent profile	Skeletal Class I w/ hyperdivergent profile	Skeletal Class II w/ hyperdivergent profile

Both regression-based methods provided numeric predictions of clinically significant skeletal angles, specifically ANB and SNMP, which were subsequently used for classification into orthodontic categories as defined in Table 1. These numeric predictions served as baseline references to evaluate and compare the predictive accuracy and clinical utility of the generative approaches discussed in later sections.

By employing these conventional regression models, we established essential performance benchmarks to contextualize the predictive capabilities and clinical advantages of advanced generative models such as diffusion-based and vision-based methods evaluated subsequently in this study.

3.2. Diffusion and latent diffusion models

A DDPM [15] is a probabilistic model designed to learn data distribution $p(x)$ by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov chain of length T. For image synthesis, the most successful models rely on a reweighted variant of the variational lower bound on $p(x)$, which mirrors denoising score matching. Such a model can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t); t = 1 \ldots T$, which are trained to predict a denoised variant of their input x_t, itself a noisy version of the input x. The corresponding objective is simplified using Eq. (1).

$$L_{DDPM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{1}$$

with t uniformly sampled from $\{1, \ldots, T\}$.

With the trained perceptual compression models consisting of encoder $\mathscr{E}$ and decoder $\mathscr{D}$, we now have access to an efficient, low-dimensional latent space in which high-frequency imperceptible details are abstracted. Compared with a high-dimensional pixel space, this space is more suitable for likelihood-based generative models because they can now (i) focus on the important semantic bits of the data, and (ii) training in a computationally efficient, lower-dimensional space. As the forward process is fixed, z_t can be efficiently obtained from $\mathscr{E}$ during training, and the samples from $p(z)$ can be decoded into the image space with a single pass through $\mathscr{D}$. Therefore, Eq. (1) can be rewritten as follows:

$$L_{LDM} = \mathbb{E}_{\mathscr{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] \tag{2}$$

with x replaced by $z = \mathscr{E}(x)$.

Diffusion models are generative models that can model the conditional distributions of the form $p(z|y)$. This can be implemented with a conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$, enabling control of the synthesis process through inputs y such as age, gender, and treatment device. A DDPM can be converted into a more

Flexible conditional image generator by augmenting its underlying U-Net backbone with a cross-attention mechanism, which is effective for learning the attention-based models of various input modalities, including temporal information. To pre-process y from various modalities in an LDM, a domain specific encoder τ_θ projects y to an interme-

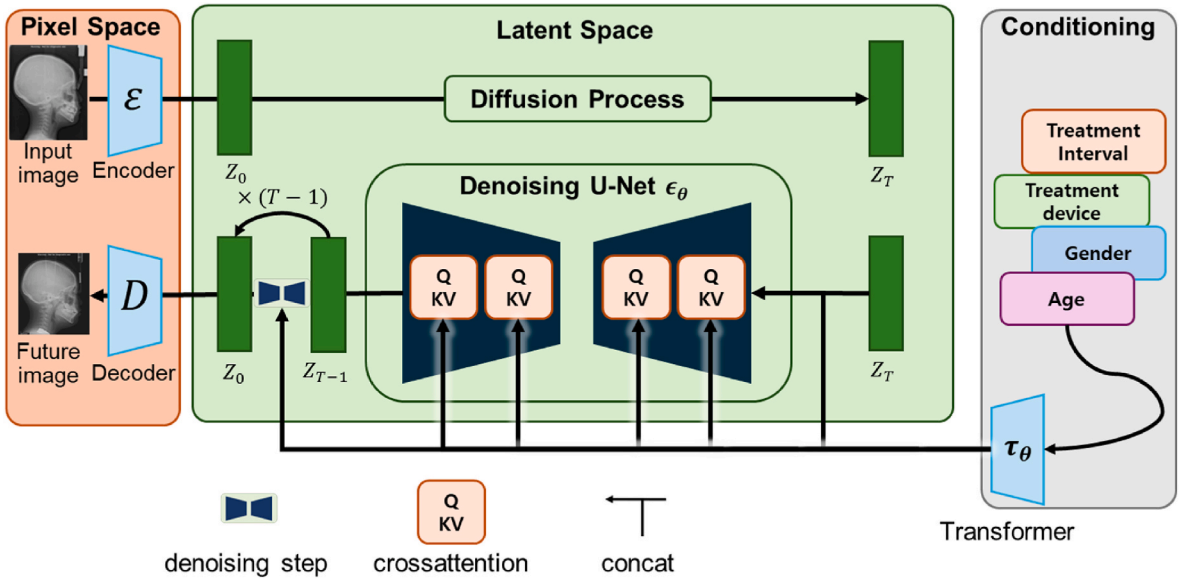

Fig. 1. Depiction of the LDM with concatenation and cross-attention mechanisms via latent space representation.

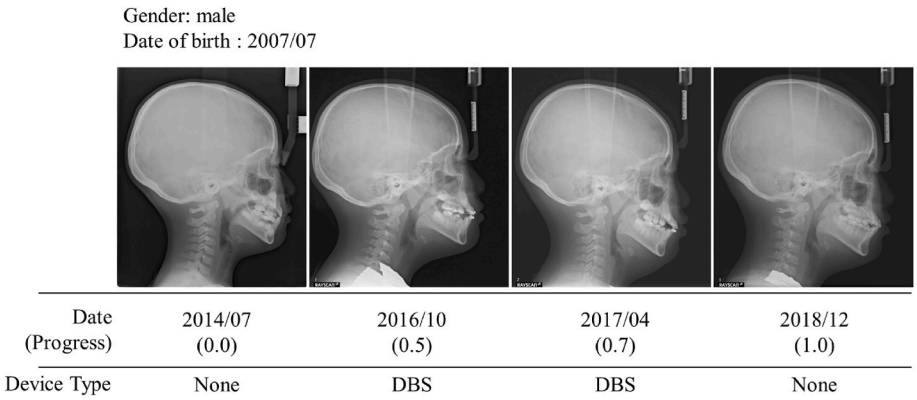

Fig. 2. Prepared data samples.

diate representation $\tau_\theta(y) \in \mathbb{R}^{M\times d_\tau}$, which is then mapped to the intermediate layers of the U-Net via a cross-attention layer implementing $\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V$, with

$$Q = W_Q^{(i)}\cdot\varphi_i(z_t), K = W_K^{(i)}\cdot\tau_\theta(y), V = W_V^{(i)}\cdot\tau_\theta(y). \tag{3}$$

Here, $\varphi_i(z_t) \in \mathbb{R}^{M\times d_\epsilon^i}$ denotes a (flattened) intermediate representation of the U-Net. A visual representation of this process is shown in Fig. 1. Based on the image-conditioning pairs, we train the conditional LDM using Eq. (4),

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t,t,\tau_\theta(y))\|_2^2\right] \tag{4}$$

3.3. Data collection and preprocessing

This study included patients who visited the Department of Orthodontics at Yonsei University Dental Hospital and underwent lateral cephalometric radiography between January 2006 and June 2022. The participants were aged 5–19 years and had a minimum of four cephalograms available. The exclusion criteria included craniofacial deformities, lesions in the craniofacial region, congenital absence of the maxillary or mandibular central incisors, and insufficient image quality for accurate landmark identification. A total of 2311 patients were identified using the Yonsei University Medical Center's SCRAP program. After applying the exclusion criteria, 14,475 cephalograms were deemed suitable for the analysis. The final dataset consisted of cephalometric X-ray images from 120 patients, each with 4–10 time-series images, resulting in an irregularly sampled dataset. Of these, 90 patients (698 images) were allocated to the training set and 30 patients (111 images) were reserved for testing. A summary of the collected datasets is presented in Fig. 2.

To ensure uniformity in terms of image size, all the X-ray images were resized to a resolution of 256 × 256 pixels. For images with aspect ratios other than 1:1, the top portions of the images were consistently cropped to maintain a standardized aspect ratio across the dataset. Because the images were X-rays, they were converted to grayscale, with pixel values normalized to lie between 0 and 1 using min-max normalization.

Each image in the dataset was labeled with specific patient attributes, including sex, age, treatment stage, type of orthodontic device used, and time information associated with the image. Sex was labeled as 0 for males and 1 for females. Age was converted into months and represented in a one-hot encoded format across 1200 bins, covering up to 100 years. Treatment progress was categorized into seven stages: initial visit (0), mid-stage of first treatment (0.5), completion of first treatment (1), growth observation (1.5), start of second treatment (2), mid-stage of second treatment (2.5), and completion of second

treatment (3). Each treatment stage is encoded across 30 bins to represent values from 0.0 to 3.0. The type of orthodontic device was labeled using a two-bit encoding system, where [0,0] represents no device; [0,1] represents devices such as RPE (Rapid Palatal Expander), Lali (Labio-Lingual appliance), and TPA (Transpalatal arch); and [1,1] represents the DBS (Direct Bonding System).

Temporal information was included as part of the input conditions, and the model used time as the predictive factor. This enabled the model to generate time-series predictions based on the progression of skeletal development, thereby improving its ability to forecast changes in a single input image.

All label conditions are concatenated and passed through an embedding layer to form a unified input for the model. This process ensured that the images, patient-specific conditions, and temporal information were fully utilized in the time-series prediction process.

3.4. Evaluation metrics

The model performance was evaluated based on a combination of image quality and clinical accuracy metrics, ensuring that the generated images were not only visually accurate but also clinically useful for orthodontic treatment planning. Using both sets of metrics, we comprehensively measured the capability of the model to generate high-fidelity images, while maintaining clinical relevance.

Three primary image quality metrics were used to evaluate the visual quality of the generated cephalometric X-ray images: MSE, SSIM, and FID. The MSE, a pixel-wise metric that calculates the average squared difference between the generated image and its corresponding ground truth (GT) image, is calculated using Eq. (5).

$$\mathrm{MSE}(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x-y)^2, \tag{5}$$

where N is the number of pixels, x is the GT pixel value, and y is the predicted pixel value.

A lower MSE value indicated that the model produced images that were more similar to the GT at the pixel level. Although MSE is a simple and commonly used metric, it may not sufficiently capture the perceptual quality, necessitating additional complementary metrics.

SSIM was used to assess the perceptual quality of the images by comparing the structural information between the generated and GT images. This metric evaluates the image similarity based on luminance, contrast, and structure, which are crucial aspects for evaluating medical images. A higher SSIM value indicates that the structural integrity of the skeletal features in the generated image closely resembles that in the original image, making it a more robust measure of image quality than MSE alone. SSIM was calculated as follows:

$$\mathrm{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)} \tag{6}$$

where μ_x and μ_y are the mean intensities of images x and y, respectively; σ_x^2 and σ_y^2 are the variances of x and y; σ_{xy} is the covariance between x and y; and C_1 and C_2 are small constants added to stabilize the division.

The FID was employed to measure the perceptual realism of the generated images by comparing the distribution of features in the generated images to those in the real images. The FID operates by extracting features from a pretrained neural network and computing the Wasserstein distance between the real and generated images in the feature space. A lower FID value suggests that the generated images are more visually realistic and exhibit characteristics similar to those of the real images, making this metric particularly important for evaluating the overall quality of the synthesized cephalometric images. The FID was calculated as follows:

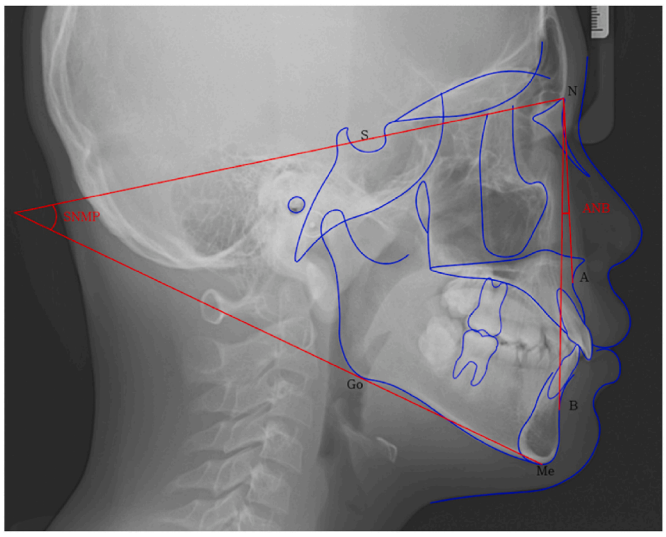

Fig. 3. Measurements of cephalometric parameters. The ANB angle was defined as the angle formed by points A (subspinale), N (nasion), and B (supramentale). It is commonly used to evaluate the anteroposterior skeletal relationship between the maxilla and mandible. Similarly, the SNMP angle refers to the angle between the sella-nasion (SN) plane and the mandibular plane, which is defined by a line connecting points Go (gonion) and Me (menton). This angle is widely used to assess vertical skeletal relationships.

$$\mathrm{FID}(x,y)=\left\|\mu_x-\mu_y\right\|^2+\mathrm{Tr}\left(C_x+C_y-2\left(C_xC_y\right)^{\frac{1}{2}}\right) \tag{7}$$

where μ_x and μ_y are the mean feature vectors of the real and generated images, respectively; C_x and C_y are the covariance matrices of the real and generated image features, respectively; and Tr denotes the trace of the matrix (the sum of its diagonal elements).

In addition to assessing image quality, we evaluated the clinical accuracy of the model predictions using classifications based on skeletal morphology. The skeletal morphology in the anteroposterior dimension was categorized into three groups: normal maxillomandibular relationships (Class I), mandibular retrognathism (Class.

II), and mandibular prognathism (Class III). Vertically, the profiles were classified as hyperdivergent, normodivergent, or hypodivergent. To ensure the applicability of the model in actual orthodontic practice, we used the ANB angle (measured between points A, N, and B) for anteroposterior classification and the SN-mandibular plane angle (SNMP) for vertical classification. The SNMP angle was determined by measuring the angle between the sella-nasion line and the mandibular plane, which is defined as the line connecting the gonion and menton. Classification thresholds were defined as follows: ANB angles greater than 5° indicates Class II, values between 2° and 5° indicate Class I, and values less than 2° indicate Class III. For vertical classification, an SNMP angle greater than 37° was classified as hyperdivergent, values between 31° and 37° as normodivergent, and values less than 31° as hypodivergent. To evaluate the model accuracy, we compared the classification results derived from the ANB and SNMP angle values between the actual patient images and the model's predicted images. These measures provide a practical framework for assessing the clinical relevance of our predictions, ensuring that the model not only delivers high-quality images but also offers actionable insights for treatment decisions. The measurement illustration is shown in Fig. 3, the classification of skeletal morphology is shown in Table 1, and the definitions of the cephalometric landmarks used are detailed in Supplement 1. These skeletal characteristics are critical for treatment planning, particularly when addressing facial aesthetics and functional outcomes.

By combining image quality metrics with clinically relevant measures, we ensured that the model's predictions were not only visually

Table 2
Numeric accuracy of the four landmark-driven regressors. Values are mean ± SD over the test set; lower is better.

Model	ANB MAE (°)	SN-MP MAE (°)
1-input SVR	2.42 ± 1.71	3.54 ± 2.05
3-input SVR	1.39 ± 1.02	2.28 ± 1.46
1-input MLP	2.73 ± 1.85	3.81 ± 2.27
3-input MLP	1.66 ± 1.18	2.55 ± 1.63

accurate, but also representative of actionable insights for clinical decision-making. This dual approach guarantees that the generated images are useful in real-world clinical settings, where accurate prediction of skeletal structures and facial profiles is essential for successful orthodontic treatment planning.

3.5. Training strategies

Three different models were trained and evaluated to predict future cephalometric images based on varying input configurations: a DDPM using three sequential input images (3-input model), a DDPM using a single input image (1-input model), and an LDM trained with a single input image.

The first approach utilizes DDPM with three sequential input images captured at different stages of treatment and used as inputs to predict future skeletal development. The inclusion of multiple images enhanced the ability of the model to learn temporal patterns of skeletal growth, thereby improving the accuracy of its predictions. However, this approach has limited clinical practicality, because it requires multiple images to be captured over time, making it unsuitable for real-time or initial treatment planning.

To address the limitations associated with the multi-image approach, we explored a second approach that uses a single-input image. The DDPM-based model was designed to predict future cephalometric changes using only the first image obtained during the initial visit. The model leverages conditioning on patient-specific attributes, such as age, sex, treatment process, and treatment device, enabling it to produce time-series predictions based on these conditions. The training process was similar to that of the multi-image model, except that the input consisted of a single image rather than a sequence.

As an extension of the second approach, the third model is an LDM with single-image input. Unlike the standard 1-input model, LDM leverages a cross-attention mechanism that allows the model to better capture the temporal dependencies inherent to time-series data, even with only one input image. This enhanced ability to incorporate and process conditional data enhances predictive accuracy, while ensuring that both the realism and clinical importance of predictions are maintained. The LDM was trained using the same denoising objective as the standard DDPM. However, by performing operations in the latent space, the model not only reduced computational complexity but also secured higher accuracy, making it more robust for time-series predictions in clinical settings.

4. Results and discussion

4.1. Accuracy evaluation using regression based models

As shown in Table 2, the three-input landmark regressors showed a clear numerical advantage over their single-input counterparts; however, even the best configuration remained outside a clinically negligible margin. For ANB, the MAE decreased from 2.8° with the 1-input SVR to

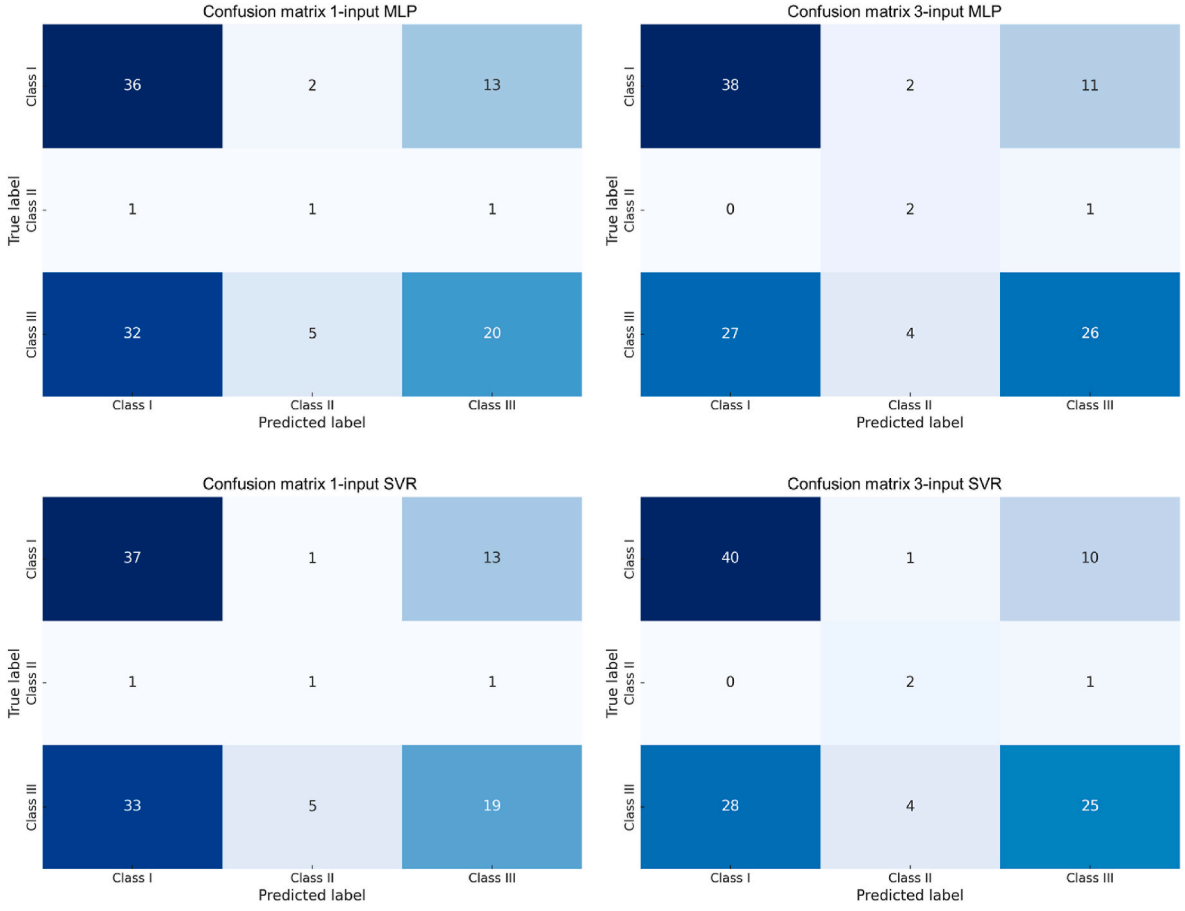

Fig. 4. Multi-class (ANB) confusion matrix of prediction models.

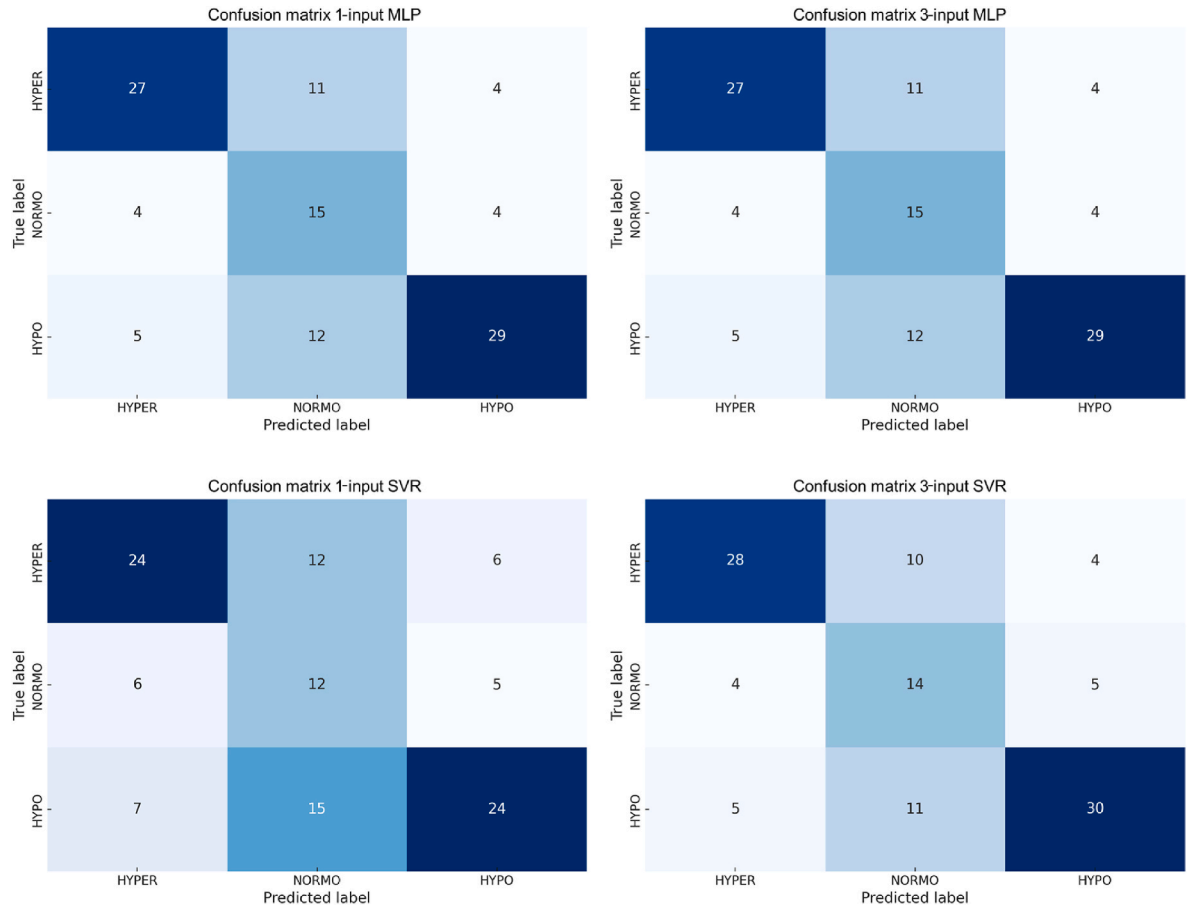

Fig. 5. Multi-class (SNMP) confusion matrix of prediction models.

1.6° with the 3-input SVR, and for SNMP, the error decreased from 4.0° to 2.8°. A similar but slightly smaller improvement was observed for the multilayer perceptron (from 3.0° to 1.9° in ANB and from 4.3° to 3.1° in SNMP). Despite these reductions, the residual error for ANB remained approximately 1.5°, and the vertical error remained close to 3°. In everyday practice, a one-degree change in ANB can shift a borderline case between Class I and Class II, while a three-degree change in SNMP can move a patient from a normodivergent to a borderline hyperdivergent status. Therefore, the present results indicate that although additional landmark snapshots improve classical regressors, the overall precision still falls short of the resolution required for confident growth forecasting from routinely acquired images. These limitations motivated the diffusion-based approach evaluated in the following sections, which seeks to deliver higher accuracy without the need for multiple follow-up radiographs.

4.2. *Clinical accuracy evaluation with regression based models*

Clinical accuracy was assessed by converting numeric predictions from the regression models into diagnostic categories defined for sagittal (ANB: Classes I, II, and III) and vertical (SNMP: hyperdivergent, normodivergent, and hypodivergent) skeletal patterns. Figs. 4 and 5 present detailed confusion matrices for visualizing the classification performance, highlighting the specific strengths and weaknesses of the models in diagnostic classification.

As shown in Fig. 4, the ANB confusion matrices revealed marked challenges for all regression-based approaches, particularly for Class III prediction accuracy. For instance, the 1-input SVR correctly identified only 19 of 57 true Class III cases, frequently misclassifying many of them as Class I. Even with additional temporal landmark inputs, the 3-input SVR showed modest improvement, correctly identifying only 25 Class III patients, indicating that over half of the clinically critical Class III cases remained inaccurately classified. The MLP models demonstrated a similar pattern of misclassification, with the 1-input MLP correctly classifying only 20 Class III patients and the 3-input MLP improving only slightly to correctly classify 26 patients. Persistent inaccuracies in distinguishing Class III cases underscore the critical weaknesses of numeric regression-based methods, particularly in cases where early diagnosis and intervention are essential for favorable orthodontic outcomes.

In the vertical dimension, the SNMP confusion matrices in Fig. 5 indicate a somewhat improved but still clinically insufficient accuracy in predicting hyperdivergent cases. The 3-input SVR, for example, accurately classified 28 of the 42 hyperdivergent patients. However, misclassifications occur frequently, with several hyperdivergent cases mistakenly classified as normodivergent. The normodivergent and hypodivergent classes also exhibited considerable misclassification, which was particularly evident in the 1-input models. Misclassification of vertical growth patterns can lead to problematic clinical decisions, as it may affect the selection of orthodontic appliances and the overall direction of treatment, ultimately influencing treatment outcomes.

The overall limited clinical accuracy of regression-based models stems from fundamental limitations inherent in landmark-based numerical approaches. First, landmarks represent only discrete spatial coordinates without encoding the complete craniofacial morphology. Second, craniofacial structures do not grow linearly but develop in diverse three-dimensional directions depending on the growth phase, making it difficult to reliably capture such changes using only a small number of landmark vectors. Although a coordinate system based on the stable structures method was employed to address this issue, some degree of error remains inevitable, as research on cephalometric

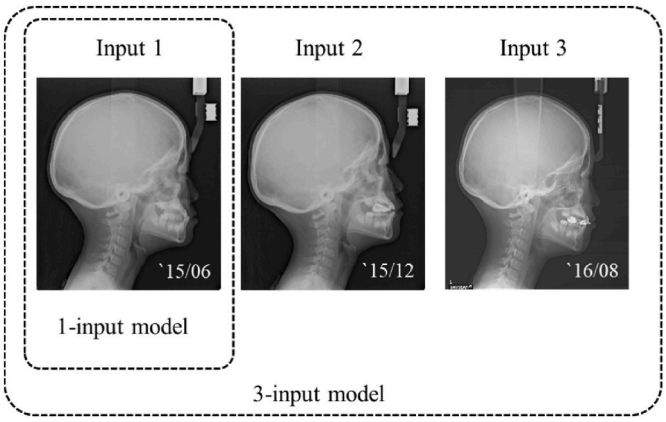

Fig. 6. Input images used for evaluation. For the 3-input model, all three images were used. For the 1-input model, only "Input 1" was used.

superimposition is still ongoing [32].

Most importantly, conventional numerical regression models inherently lack visual representation capabilities. Unlike diffusion-based image synthesis methods, which provide clinicians with visual forecasts that can be intuitively evaluated, measured, and verified, regression-based numerical predictions do not provide image-level visualizations. Consequently, clinicians have limited opportunities to interpret or validate predictions visually, significantly constraining the clinical utility and trustworthiness of landmark-only numerical regression predictions. This key limitation, along with demonstrated clinical inaccuracies, underscores the motivation for adopting diffusion-based image synthesis methods that address these fundamental limitations and offer enhanced clinical applicability.

4.3. Vision-based accuracy evaluation with image synthesis models

We evaluated the visual quality of the cephalometric images generated by four different models: the 3-input DDPM, 1-input DDPM, 1-input LDM, and 1-input ControlNet [33]. Although ControlNet was not previously discussed in earlier sections, it is introduced here as an additional generative model to further validate predictive accuracy using limited input data. The evaluation was based on quantitative metrics including MSE, SSIM, and FID, alongside qualitative visual inspections of anatomical accuracy. Fig. 6 provides the sample input images used for the evaluation, and visual comparisons of the model predictions are presented in Fig. 6 (entire images) and Fig. 7 (region-of-interest [ROI] images).

Table 3 summarizes the quantitative evaluation results across entire images. The 3-input DDPM, benefiting from sequential temporal inputs, achieved the lowest MSE (0.0102) and highest SSIM (0.596), indicating superior pixel-level accuracy and structural integrity. However, despite using only one image, the 1-input LDM and the.

Newly evaluated 1-input ControlNet demonstrated competitive

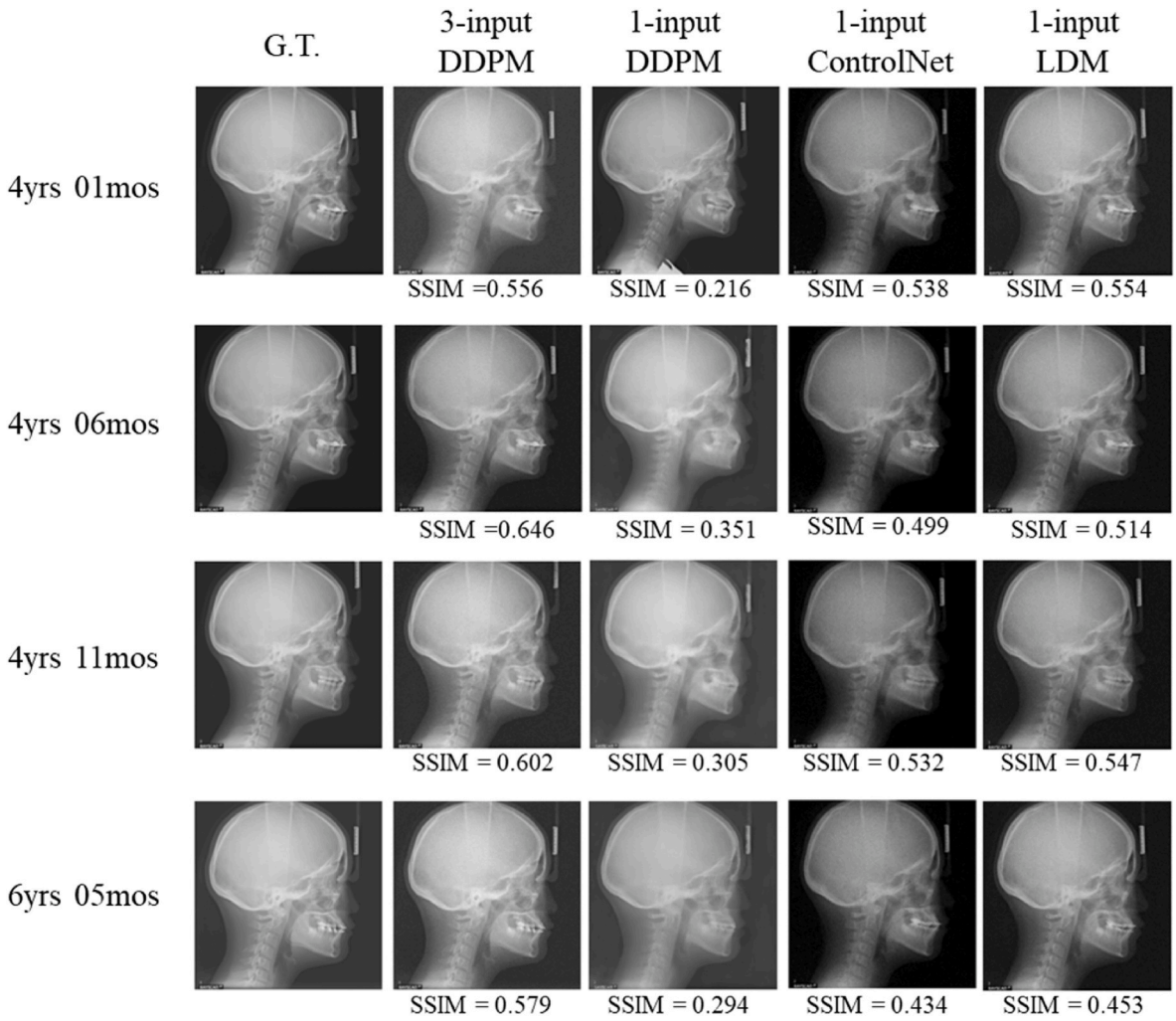

Fig. 7. Qualitative plot of prediction results for 3-input model and 1-input models.

Table 3
Quantitative results of 3-input models and 1-input models.

Metric	3-input DDPM	1-input DDPM	1-input ControlNet	1-input LDM
MSE (↓)	0.0102	0.0246	0.0141	0.0138
SSIM (↑)	0.596	0.301	0.468	0.471
FID (↓)	91.32	143.75	83.46	81.94

Table 4
Quantitative results of 3-input models and 1-input models for ROIs.

Metric	3-input DDPM	1-input DDPM	1-input ControlNet	1-input LDM
MSE (↓)	0.00419	0.0112	0.0125	0.0134
SSIM (↑)	0.647	0.141	0.464	0.472
FID (↓)	67.17	230.75	85.96	82.94

performance. Specifically, the 1-input ControlNet achieved an MSE of 0.0141 and SSIM of 0.468, closely matching the 1-input LDM (MSE 0.0138, SSIM 0.471), both significantly outperforming the baseline 1-input DDPM (MSE 0.0246, SSIM 0.301). Regarding perceptual realism measured by FID scores, the 1-input ControlNet exhibited remarkable results (FID 83.46), slightly surpassing the 1-input LDM (FID 81.94), and significantly outperforming both the 3-input DDPM (FID 91.32) and 1-input DDPM (FID 143.75). These metrics underscore the capability of both ControlNet and LDM to achieve high visual realism and structural fidelity despite limited input data.

Detailed ROI analyses (Table 4 and Fig. 8) confirmed these trends. In terms of MSE, the 3-input DDPM again showed superior accuracy (0.00419), reflecting the advantages of having multiple time points. The 1-input ControlNet and LDM yielded similar MSE values (0.0125 and 0.0134, respectively), clearly outperforming the 1-input DDPM (0.0112). The SSIM values reinforced these findings, with the 3-input DDPM achieving the highest structural similarity (0.647), followed by ControlNet (0.464) and LDM (0.472), both significantly exceeding the 1-input DDPM (0.141). For perceptual realism measured by FID within ROIs, the 3-input DDPM maintained superiority (FID 67.17), but notably, the 1-input ControlNet (FID 85.96) closely matched the 1-input LDM (FID 82.94), and both significantly outperformed the 1-input DDPM (FID 230.75).

Qualitative visual assessments (Figs. 7 and 8) further supported these quantitative evaluations. The 1-input ControlNet, leveraging contrastive learning through content and style representation, clearly enhanced anatomical accuracy and reduced visual artifacts compared to the 1-input DDPM. This resulted in predictions that were structurally coherent and visually realistic, closely resembling clinical expectations.

Overall, these evaluations indicate that the 1-input ControlNet significantly improves predictive accuracy and visual realism compared to the baseline 1-input DDPM. Although the 3-input DDPM maintained slight numerical advantages in pixel-level metrics, the performance of the 1-input ControlNet and LDM demonstrated remarkable potential in clinical scenarios requiring minimal imaging. ControlNet, in particular, offers substantial clinical practicality due to its combination of accuracy, visual quality, and computational efficiency, highlighting its value in personalized orthodontic treatment planning with reduced imaging

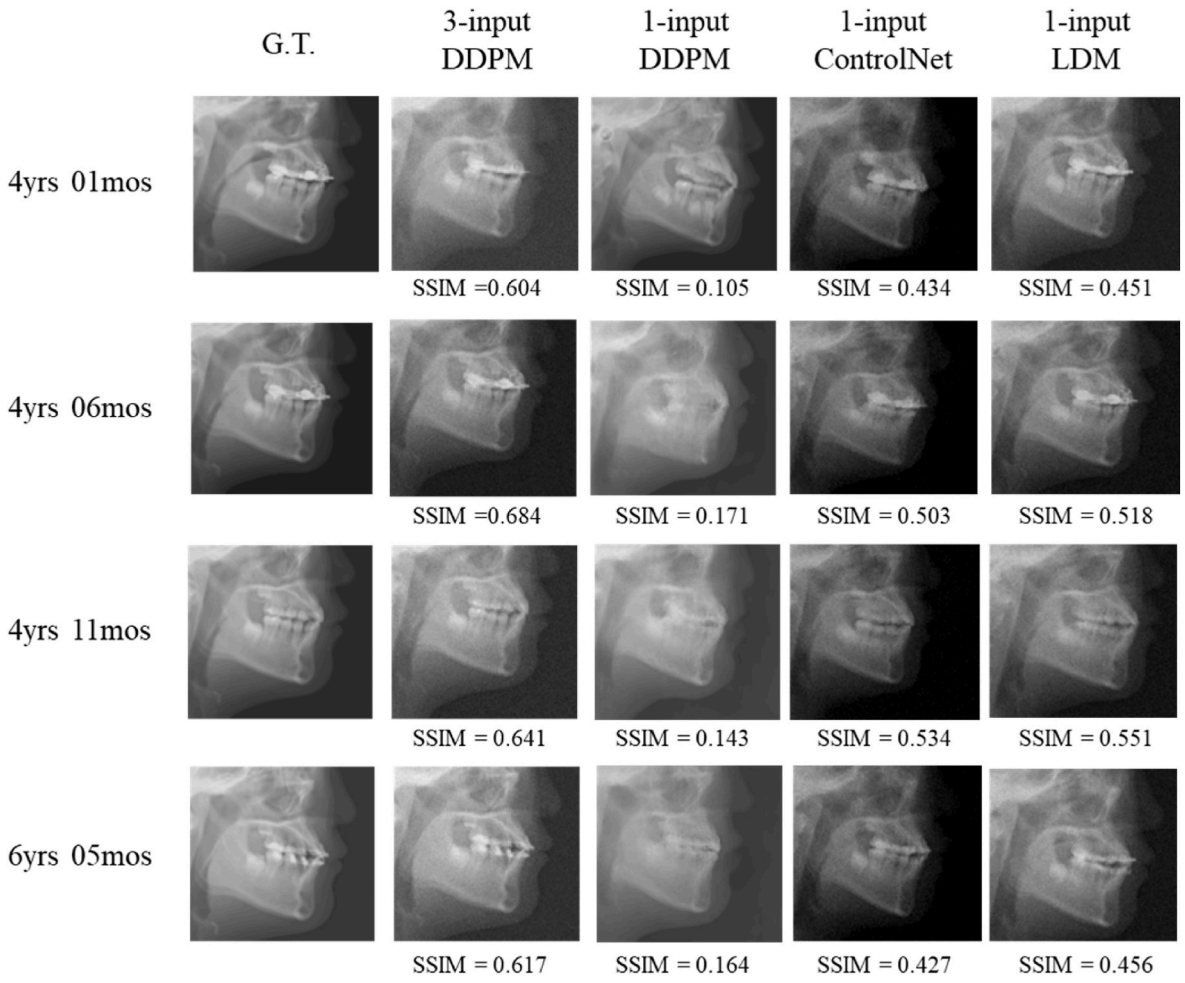

Fig. 8. Qualitative plot of prediction results of 3-input model and 1-input models for ROIs.

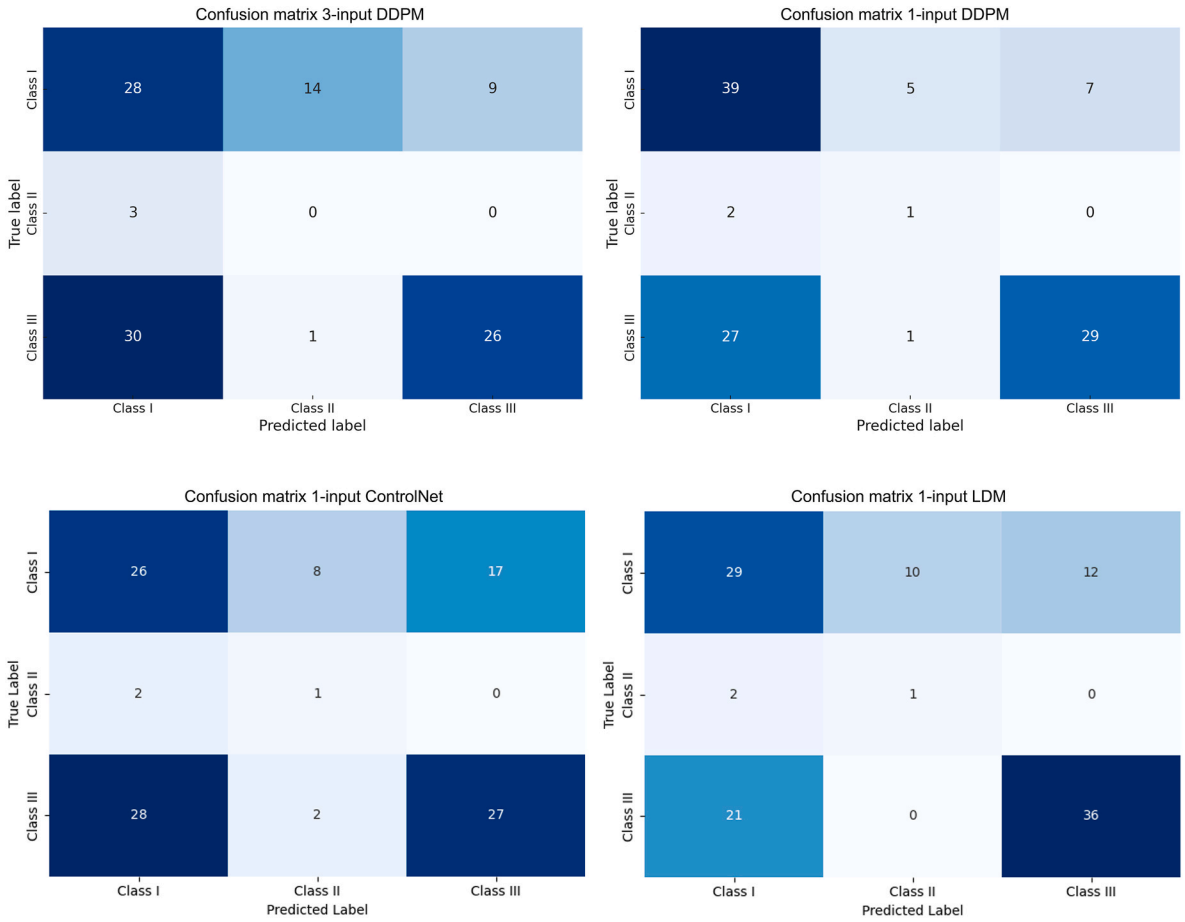

Fig. 9. Multi-class (ANB) confusion matrix of prediction models.

Table 5
Prediction accuracy of 3-input model and 1-input models.

	3-input DDPM	1-input DDPM	1-input ControlNet	1-input LDM
Sagittal relationship (ANB angle)	48.6 %	62.2 %	48.6 %	59.5 %
Vertical pattern (SNMP angle)	58.6 %	63.1 %	61.2 %	64.9 %

frequency.

4.4. Clinical accuracy evaluation with DDPM based models

A multiclass confusion matrix presented the clinical accuracies by comparing the ANB values between the actual patient images and the predicted images from four different models (Fig. 9). The prediction accuracies of the individual models were as follows: the 3-input DDPM model and 1-input ControlNet model had the lowest accuracy at 48.6 %, the 1-input DDPM model had the highest accuracy at 62.2 %, and the 1-input LDM had an accuracy of 59.5 %, as shown in Table 5. The 1-input LDM demonstrated a significantly better performance than the 3-input DDPM, and showed results similar to those of the 1-input DDPM, indicating that it is possible to predict skeletal growth and changes from a single initial image. Therefore, using an 1-input LDM that does not require three time-series data inputs, such as the 3-input DDPM, is also clinically effective.

All four models failed to exceed a 70 % accuracy rate. The inherent challenges of clinical studies may partially explain these results. This retrospective study used a dataset of patients who had undergone treatment. Collecting a longitudinal dataset from growing patients without therapeutic intervention is nearly impossible, which means that the models may face unique challenges compared with predicting natural growth patterns. Additionally, the high proportion of Class III malocclusion cases and their treatment characteristics may have influenced the outcomes. In the prediction results, cases of skeletal Class III malocclusion were frequently misclassified, with predictions nearly evenly split between Class I and Class III malocclusions. Treatment of Class.

III malocclusion often involves anterior maxillary traction, which can significantly alter ANB values within a year. Consequently, Class III malocclusion cases are frequently reclassified as Class I or Class II malocclusions within a short timeframe [34]. Conversely, some Class III cases progress to severe conditions requiring orthognathic surgery due to continued mandibular growth, which is challenging to predict [35,36]. This variability and complexity may contribute to inconsistencies in the prediction process.

The prediction results for the vertical pattern exhibit a similar trend in the multiclass confusion matrix (Fig. 10). Amongst the models, the 3-input DDPM model showed the lowest prediction accuracy (58.6 %), followed by the 1-input ControlNet model (61.2 %) and the 1-input DDPM model (63.1 %). 1-input LDM achieved the highest accuracy at 64.9 % (Table 5). Similar to the predictions for the anteroposterior relationship, the overall prediction accuracy for the vertical relationship was relatively low. Variability introduced by natural growth and treatment-induced changes likely affected the prediction accuracy. Regarding natural growth, the SNMP angle decreases during the growth phase [37,38], which may explain why the models predominantly predicted a hypodivergent pattern. However, this tendency appears to be somewhat overpredicted compared to the actual patient data.

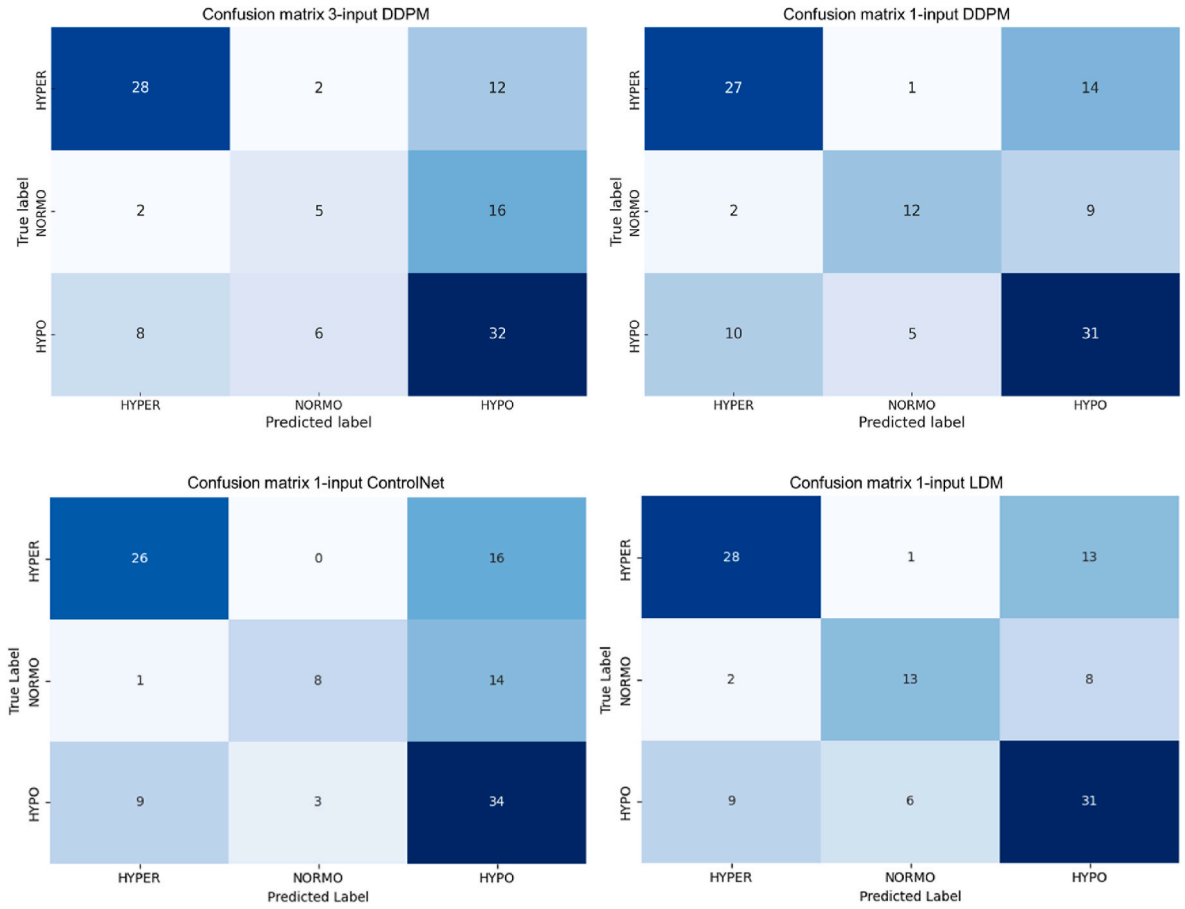

Fig. 10. Multi-class (SNMP) confusion matrix of prediction models.

Furthermore, the initial craniofacial morphology shows only a weak correlation with the direction of morphological changes during growth [39], and growth patterns during adolescence can differ significantly from those observed during childhood [35]. These factors likely contribute to the variability in predictions related to natural growth. In addition to natural growth, treatment-induced changes are also likely to play a significant role. The dataset included a large proportion of cases with noticeable skeletal changes occurring early in the treatment process, which may have introduced additional variability into the model predictions. These early treatment-related changes, combined with the inherent variability of natural growth, likely pose challenges to the artificial intelligence models in achieving higher prediction accuracy.

Despite the challenges in prediction based on initial morphology, our study demonstrated that 1-input LDM could predict skeletal morphology at a specific time point during growth using initial radiographs. This result has greater clinical significance than predicting natural growth alone because it incorporates skeletal changes associated with orthodontic treatment. It can greatly aid in initial consultations by predicting and visualizing skeletal changes resulting from orthodontic treatment, thereby providing a rationale for addressing skeletal issues. Furthermore, it can facilitate the development of treatment plans to address dentoalveolar issues. However, improving the prediction accuracy requires further studies with a larger dataset. Additionally, enhancing the image resolution to differentiate dental structures could allow for further cephalometric measurements and analyses, thereby increasing the diagnostic value of the model.

4.5. Summary of clinical and vision-based accuracy comparisons across entire models

Fig. 11 provides a comprehensive summary and direct comparison of the evaluated predictive models, clearly illustrating clinical accuracy in terms of ANB and SNMP classification accuracy, alongside vision-based accuracy represented by SSIM and FID scores. Fig. 11. (a) compares clinical accuracy across regression-based models including SVR and MLP, vision-based models such as ControlNet, and diffusion-based models DDPM and LDM. The results highlight that diffusion-based approaches generally outperform traditional methods in clinical classification accuracy. Notably, the 1-input LDM and 1-input DDPM demonstrate comparable clinical accuracy among all single-image input models.

Fig. 11. (b) further evaluates image quality through vision-based metrics, specifically SSIM and FID. Among single-image models, the 1-input LDM clearly achieves the highest structural similarity and lowest perceptual difference, significantly outperforming the 1-input DDPM and ControlNet. While the 3-input DDPM model exhibited slightly superior image quality overall, the balanced performance of the 1-input LDM makes it highly advantageous for single-image predictions.

Considering both clinical and vision-based evaluations, these analyses emphasize that the 1-input LDM provides an optimal balance of high clinical accuracy and superior image quality from minimal input data. Thus, the 1-input LDM emerges as the most practical and effective predictive model among single-image options, particularly suitable for clinical orthodontic applications.

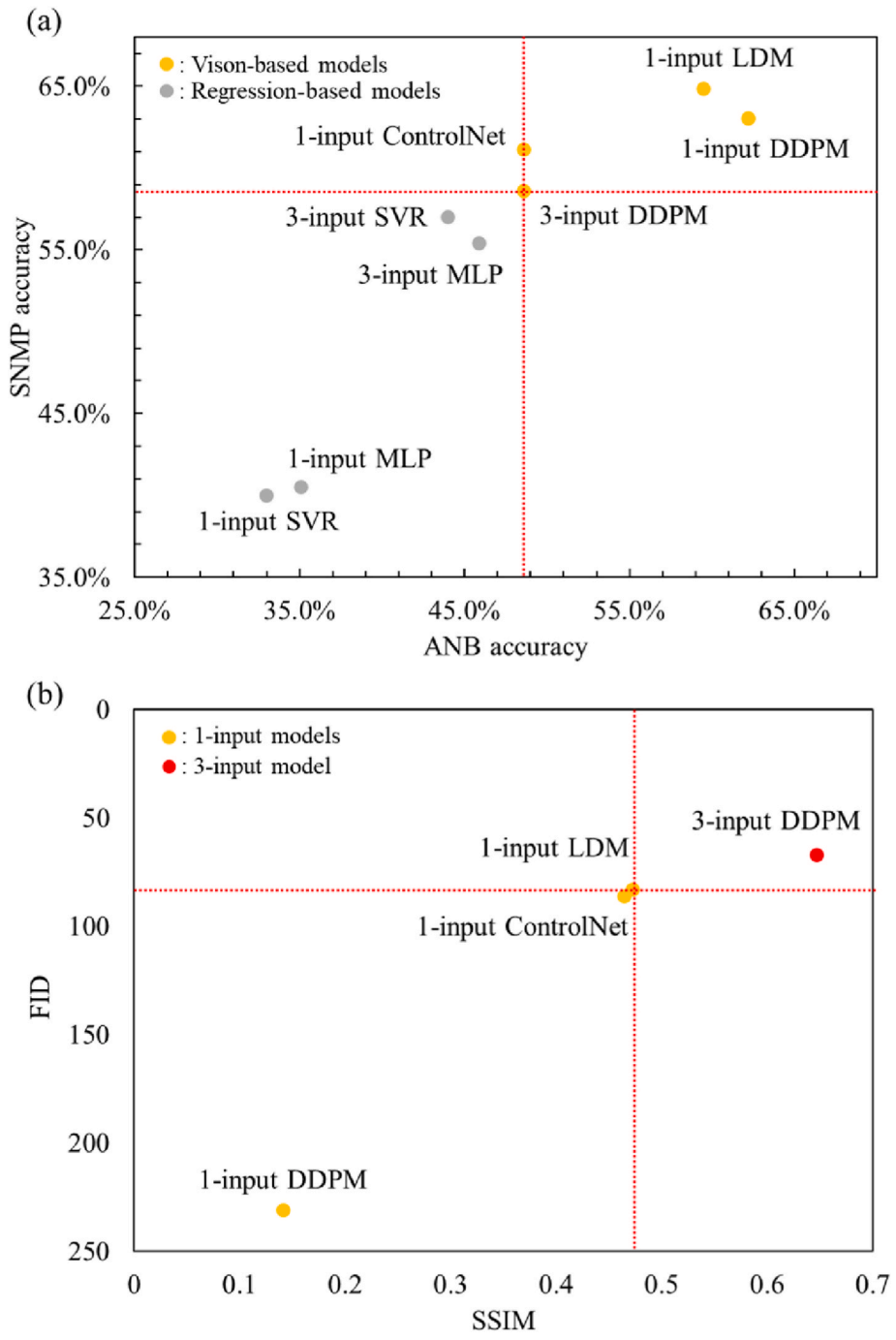

Fig. 11. (a) Clinical accuracy comparison for ANB versus SNMP classification accuracy across entire models. (b) Vision-based accuracy comparison of SSIM versus FID scores for vision-based models.

5. Conclusion

This study explored the utility of advanced generative models, specifically the 3-input DDPM, 1-input DDPM, 1-input LDM, and 1-input ControlNet, for predicting skeletal changes during orthodontic treatment from cephalometric images. Our evaluation combined quantitative image quality metrics such as MSE, SSIM, and FID with qualitative assessments and clinical accuracy to demonstrate that these models effectively predict future skeletal morphology with varying degrees of accuracy and practicality.

Quantitatively, the 3-input DDPM model achieved the best numerical accuracy, demonstrating the lowest overall MSE at 0.0102 and highest SSIM at 0.596. These results highlight the advantage of sequential temporal information in capturing detailed anatomical changes. Despite using only a single image, the 1-input LDM and the newly evaluated 1-input ControlNet showed impressive performance, closely matching the accuracy of the multi-image DDPM. The 1-input ControlNet model exhibited significant improvements over the baseline 1-input DDPM, achieving a substantially lower MSE of 0.0141 compared to 0.0246, a higher SSIM of 0.468 compared to 0.301, and a notably superior FID score of 83.46 compared to 143.75. These improvements confirm that the 1-input ControlNet, enhanced by contrastive learning mechanisms, successfully improved structural coherence and perceptual realism in generated images, making it highly practical for clinical applications with minimal imaging requirements.

Clinical accuracy evaluations revealed notable strengths and

challenges for each model. Although the 3-input DDPM provided the most precise pixel-level predictions, the clinical practicality of routinely capturing multiple sequential images is limited. Conversely, both the 1-input LDM and ControlNet effectively balanced predictive accuracy and clinical convenience, enabling reliable predictions from a single image. The ControlNet model demonstrated substantial improvements in predicting complex anatomical scenarios, reflecting its effectiveness in integrating patient-specific conditions through contrastive learning.

Despite these promising outcomes, the study has certain limitations that present opportunities for future research. Predictive accuracy, particularly in challenging Class III and hyperdivergent cases, still requires improvement. Additionally, while the 1-input models are clinically practical, further optimization and validation on larger and more diverse datasets are necessary to enhance generalizability and robustness.

In conclusion, our findings highlight the potential of 1-input ControlNet and LDM, as powerful tools in orthodontic treatment planning. These models significantly reduce patient exposure to radiation and imaging frequency while maintaining high predictive accuracy and clinical applicability. Further development and validation of these models could significantly improve clinical workflows, offering orthodontists precise, visually interpretable predictions that inform personalized and proactive treatment strategies.

CRediT authorship contribution statement

Soon Wook Kwon: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Jung Ki Moon:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Seung-Cheol Song:** Data curation. **Jung-Yul Cha:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Young Woo Kim:** Writing – review & editing, Visualization. **Yoon Jeong Choi:** Data curation, Conceptualization, Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Formal analysis. **Joon Sang Lee:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Ethics in publishing statement

I testify on behalf of all co-authors that our article submitted followed ethical principles in publishing.

All authors agree that:

This research presents an accurate account of the work performed, all data presented are accurate and methodologies detailed enough to permit others to replicate the work.

This manuscript represents entirely original works and or if work and/or words of others have been used, that this has been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

That generative AI and AI-assisted technologies have not been utilized in the writing process or if used, disclosed in the manuscript the use of AI and AI-assisted technologies and a statement will appear in the published work.

That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research design where such use must be described in a reproducible manner in the methods section.

All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Starting growth Technological R&D Program (TIPS Program, (No. RS-2024-00441761)) funded by the Ministry of SMEs and Startups(MSS, Korea), the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (RS-2022-NR070832), and the Yonsei University College of Dentistry Fund (6-2022-0135).

References

[1] A.R. Durão, et al., Validity of 2D lateral cephalometry in orthodontics: a systematic review, Prog. Orthod. 14 (2013) 1–11.
[2] A. Gupta, On imaging modalities for cephalometric analysis: a review, Multimed. Tool. Appl. 82 (24) (2023) 36837–36858.
[3] N.M. Helal, O.A. Basri, H.A. Baeshen, Significance of cephalometric radiograph in orthodontic treatment plan decision, J. Contemp. Dent. Pract. 20 (7) (2019) 789–7793.
[4] N. Kazimierczak, et al., AI in orthodontics: revolutionizing diagnostics and treatment planning—a comprehensive review, J. Clin. Med. 13 (2) (2024) 344.
[5] P. Pittayapat, et al., Three-dimensional cephalometric analysis in orthodontics: a systematic review, Orthod. Craniofac. Res. 17 (2) (2014) 69–91.
[6] A. Kebaili, J. Lapuyade-Lahorgue, S. Ruan, Deep learning approaches for data augmentation in medical imaging: a review, J. Imag. 9 (4) (2023) 81.
[7] A. Kushwaha, et al., Rapid training data creation by synthesizing medical images for classification and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.
[8] J. Niemeijer, et al., TSynD: targeted synthetic data generation for enhanced medical image classification: leveraging epistemic uncertainty to improve model performance, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer, 2024.
[9] V. Thambawita, et al., SinGAN-Seg: synthetic training data generation for medical image segmentation, PLoS One 17 (5) (2022) e0267976.
[10] S.H. Kang, et al., 3D cephalometric landmark detection by multiple stage deep reinforcement learning, Sci. Rep. 11 (1) (2021) 17509.
[11] A. Marya, et al., Development and validation of predictive models for skeletal malocclusion classification using airway and cephalometric landmarks, BMC Oral Health 24 (1) (2024) 1064.
[12] F. Schwendicke, et al., Deep learning for cephalometric landmark detection: systematic review and meta-analysis, Clin. Oral Invest. 25 (7) (2021) 4299–4309.
[13] M. Serafin, et al., Accuracy of automated 3D cephalometric landmarks by deep learning algorithms: systematic review and meta-analysis, La radiologia medica 128 (5) (2023) 544–555.
[14] Y. Song, et al., Automatic cephalometric landmark detection on X-ray images using a deep-learning method, Appl. Sci. 10 (7) (2020) 2547.
[15] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[16] K. Rasul, et al., Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting, in: International Conference on Machine Learning, PMLR, 2021.
[17] H. Wen, et al., Diffstg: probabilistic spatio-temporal graph forecasting with denoising diffusion models, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, 2023.
[18] R. Azad, et al., Bi-directional ConvLSTM U-Net with densley connected convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[19] P. Desai, et al., Next frame prediction using ConvLSTM, in: Journal of Physics: Conference Series, IOP Publishing, 2022.
[20] S. Mukherjee, et al., Predicting video-frames using encoder-convlstm combination, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019.
[21] G. Zakhar, et al., Prediction of pubertal mandibular growth in males with class II malocclusion by utilizing machine learning, Diagnostics 13 (16) (2023) 2713.
[22] T. Wood, et al., Prediction of the post-pubertal mandibular length and Y axis of growth by using various machine learning techniques: a retrospective longitudinal study, Diagnostics 13 (9) (2023) 1553.
[23] M. Parrish, et al., Short-and long-term prediction of the post-pubertal mandibular length and Y-Axis in females utilizing machine learning, Diagnostics 13 (17) (2023) 2729.
[24] S. Kaźmierczak, et al., Prediction of the facial growth direction with machine learning methods, arXiv preprint arXiv:2106.10464 (2021).
[25] J. Ho, et al., Video diffusion models, Adv. Neural Inf. Process. Syst. 35 (2022) 8633–8646.
[26] Y. Tashiro, et al., Csdi: conditional score-based diffusion models for probabilistic time series imputation, Adv. Neural Inf. Process. Syst. 34 (2021) 24804–24816.

[27] W.H. Pinaya, et al., Brain imaging generation with latent diffusion models, in: MICCAI Workshop on Deep Generative Models, Springer, 2022.
[28] H. Tao, Erasing-inpainting-based data augmentation using denoising diffusion probabilistic models with limited samples for generalized surface defect inspection, Mech. Syst. Signal Process. 208 (2024) 111082.
[29] D. Guo, et al., Towards better cephalometric landmark detection with diffusion data generation, IEEE Trans. Med. Imag. 44 (7) (2025) 2784–2794.
[30] I.-H. Kim, et al., Predicting orthognathic surgery results as postoperative lateral cephalograms using graph neural networks and diffusion models, Nat. Commun. 16 (1) (2025) 2586.
[31] R. Di Via, F. Odone, V.P. Pastore, Self-supervised pre-training with diffusion model for few-shot landmark detection in x-ray images, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025.
[32] N. Agrawal, et al., Cephalometric superimposition in orthodontics-A review, IP Indian J. Orthodontics Dentofacial Res. 8 (1) (2022) 1–6.
[33] Z. Wang, et al., A content-style control network with style contrastive learning for underwater image enhancement, Multimed. Syst. 31 (1) (2025) 1–13.
[34] L. Dermaut, C. Aelbers, Orthopedics in orthodontics: fiction or reality. A review of the literature—part II, Am. J. Orthod. Dentofacial Orthop. 110 (6) (1996) 667–671.
[35] B.A. Chvatal, et al., Development and testing of multilevel models for longitudinal craniofacial growth prediction, Am. J. Orthod. Dentofacial Orthop. 128 (1) (2005) 45–56.
[36] Y.J. Choi, et al., Prediction of long-term success of orthopedic treatment in skeletal Class III malocclusions, Am. J. Orthod. Dentofacial Orthop. 152 (2) (2017) 193–203.
[37] S.S. Yoon, C.-H. Chung, Comparison of craniofacial growth of untreated Class I and Class II girls from ages 9 to 18 years: a longitudinal study, Am. J. Orthod. Dentofacial Orthop. 147 (2) (2015) 190–196.
[38] A. Bjo, V. Skieller, Facial development and tooth eruption: an implant study at the age of puberty, Am. J. Orthod. 62 (4) (1972) 339–383.
[39] A. Katsadouris, D.J. Halazonetis, Geometric morphometric analysis of craniofacial growth between the ages of 12 and 14 in normal humans, EJO (Eur. J. Orthod.) 39 (4) (2017) 386–394.