2026 International Conference on Automatic Face and Gesture Recognition (FG)

StyleDiT: A Unified Framework for Diverse Child and Partner Faces Synthesis with Style Latent Diffusion Transformer

Pin-Yen Chiu* Dai-Jie Wu* Po-Hsun Chu Chia-Hsuan Hsu Hsiang-Chen Chiu
Chih-Yu Wang Jun-Cheng Chen
Research Center for Information Technology Innovation, Academia Sinica

***Abstract*— Kinship face synthesis is challenging due to the scarcity and low quality of available kinship data. Existing methods struggle to balance diversity and fidelity of generated faces while precisely controlling facial attributes such as age and gender. To address these issues, we propose the Style Latent Diffusion Transformer (StyleDiT), a novel framework that integrates the strengths of StyleGAN with the diffusion model to generate diverse and high-quality kinship faces. Our conditional diffusion model mainly focuses on modeling the complex kinship distribution to allow sampling a StyleGAN latent aligned with the kinship relationship of conditioning images. The final face is then synthesized through the pretrained StyleGAN generator. The design not only significantly reduces the training complexity but also guarantees the generation diversity and quality. To further enhance control, we introduce Relational Trait Guidance (RTG), which enables independent modulation of parental influence and fine-grained trade-offs between diversity and fidelity. Meanwhile, our framework also provides precise attribute control by leveraging the rich facial priors of StyleGAN. Furthermore, we extend the application to an insufficiently explored domain: partner face prediction, using a child's image and one parent's image within the same framework. Extensive experiments demonstrate that our StyleDiT striking the best balance between generating diverse and high-fidelity kinship faces over existing baseline methods.**

I. INTRODUCTION

Recent advances in computational facial analysis have improved our understanding of parent-child visual relationships, notably in kinship verification [40], [53] and genetic studies [2], [9]. Meanwhile, progress in face synthesis has sparked interest in high-fidelity kinship face generation, with potential applications in kinship verification and predicting the possible appearances of long-lost family members. Additionally, synthetic data is being explored for training face recognition models [45], [29], [34], [5] while addressing ethical and legal concerns. Enhancing kinship face generation may help mitigate data scarcity, improving kinship verification and recognition reliability.

Early research on kinship face synthesis [36], [19], [47] approached the task as an image-to-image translation problem, learning direct parent-to-child mappings from limited, low-quality kinship data. This often resulted in low-resolution blurry images with little variation. In contrast, several works [56], [33], [8], [31], [41] extracted genetic features from parental face images to generate more faithful child faces. Recent methods [31], [8], [33], [41] leveraged the rich semantic information in the latent space of the pretrained StyleGAN [1], [27], [25] to produce high-quality child faces. StyleGAN's smooth and disentangled latent space allows for seamless fusion of parental facial features through latent interpolation. However, these approaches still struggle to strike a balance between both high diversity and fidelity in the generated faces while precisely controlling facial attributes.

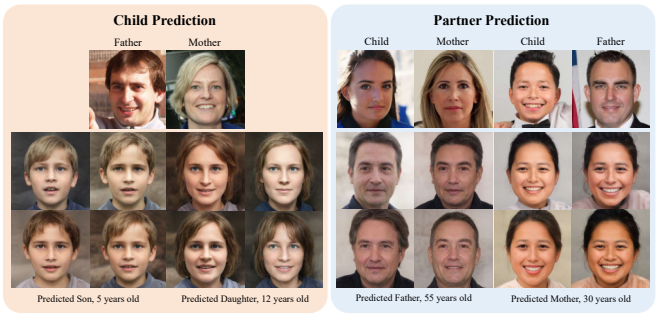


Fig. 1. StyleDiT demonstrates dual capabilities in kinship face synthesis: generating child faces from parental images and predicting partner faces given a child and one parent. For child prediction, it produces diverse and high-quality child faces across age and gender variations. In partner prediction, it effectively synthesizes realistic and varied partner faces, highlighting its flexibility across both tasks.

Moreover, the control of diversity is essential in kinship face generation since child faces generated from a parent pair can vary widely in resemblance, mirroring the natural variation seen among siblings in a family. Thus, a kinship generation model should produce diverse child faces for a given parental pair while allowing users to control the degree of resemblance to the parents. Furthermore, an innovative and potentially valuable task of partner face prediction has received limited attention. As shown in Fig. 2 and discussed in Appendix B.1, a simple linear operation in the StyleGAN latent space between a child and one parent fails to produce plausible results. This underscores the complexity of kinship distributions and highlights the need for a learning-based approach to address this challenging task.

Motivated by recent advances that combine the strengths of StyleGAN and diffusion models [38], [32], [18], [10], and recognizing the need for precise control over age and gender in kinship face synthesis, we propose Style Latent

* denotes equal contribution.
Code implementation is available at https://github.com/aiiu-lab/StyleDiT.
Acknowledgments. This research is supported by the National Science and Technology Council (NSTC), Taiwan under Grants 114-2221E-001-016, 113-2634-F-002-008, 114-2634-F-001-001-MBK, 114-2221-E-001-016, 114-2634-F-002-004, 111-2628-E-001-002-MY3, 114-2221-E-001-017-MY2, and by Academia Sinica under Grants AS-IAIA-114-M10 and AS-KPQ-112-NETZ-10-A.

979-8-3315-7231-0/26/$31.00 ©2026 IEEE

Diffusion Transformer (StyleDiT) to address key challenges in this domain. StyleDiT synergizes the fine-grained, continuous attribute control of StyleGAN's style latent space with the powerful generative capabilities of diffusion models [22], [49], which are well-suited for modeling complex distributions such as those underlying kinship relationships. In this framework, StyleGAN handles the final face generation, while our conditional diffusion model samples a StyleGAN latent that aligns with the characteristics of the conditioning images. This design ensures high diversity and fidelity while significantly reducing training complexity. As compared to Arc2Face [37], StyleDiT requires only hundreds of thousands of images and converges much faster than standard diffusion models. To further improve conditional generation, we introduce Relational Trait Guidance (RTG), an advanced Classifier-Free Guidance [23] tailored for kinship face synthesis. RTG enables independent control over each conditioning input (e.g., each parent's facial image), allowing the model to better align the generated output with the provided conditions. This flexible guidance mechanism not only improves resemblance to the conditioning inputs but also empowers users to effectively balance fidelity and diversity according to specific requirements. Meanwhile, our framework offers fine-grained attribute manipulation by leveraging the rich facial priors of StyleGAN. Furthermore, we extend kinship face synthesis to a largely unexplored and challenging task: predicting a partner's facial features given a child's image and one parent's image. This innovative task setting holds significant potential in forensic science for reconstructing the appearances of missing individuals, as well as in social and genetic research for studying inherited facial traits. Our main contributions are summarized as follows:

- We propose StyleDiT, the first unified framework for generating diverse, high-fidelity child and partner faces with precise control over age and gender attributes.
- Our Relational Trait Guidance allows independent control of each influencing factor, enabling users to balance diversity and fidelity in the generated images effectively.
- Extensive evaluation and user study show that StyleDiT strikes an excellent balance between diversity and fidelity, outperforming state-of-the-art methods and highlighting its effectiveness and innovation.

II. RELATED WORK

Deep Image Generation and Manipulation. Recent advancements in image generation and manipulation have been driven by innovative generative models. Generative Adversarial Networks (GANs) [20], particularly in frameworks like StyleGAN [26], [27], [25], enable precise manipulation and adjustment within high-dimensional latent spaces via GAN inversion, facilitating detailed edits of real images [46], [52]. Additionally, diffusion models [22], [49] have emerged as powerful tools, progressively refining image quality through iterative processes. Moreover, a promising trend is emerging in frameworks that combine the strengths of StyleGAN and diffusion models [38], [32], [18], [10]. These approaches

Fig. 2. **Failure of Linear Operation in Partner Face Prediction.** Simple linear operations that estimate a partner's style latent code by assuming a linear parent–child relationship, i.e., $S_{M/F} = 2S_C - S_{F/M}$ derived from $S_C = \frac{1}{2}(S_F + S_M)$ in the StyleGAN latent space, fail to produce realistic partner faces. Here, S_C, S_F, and S_M denote the child, father, and mother style latent representations, respectively. This highlights the need for a learning-based approach to model the complex kinship distributions.

synergize the disentangled and fine-grained facial attribute control of StyleGAN's latent space with the versatility of text-driven image generation in diffusion models. Our StyleDiT framework advances this concept by integrating the precise facial attribute manipulation of StyleGAN's latent space with the diffusion model's ability to capture the underlying distribution of kinship relationships. Specifically, we incorporate the S space [51] of StyleGAN into a one-dimensional transformer-based diffusion model [17], [39], moving beyond the conventional focus on predicting latent codes in the W or $W+$ spaces [42], [54].

Kinship Face Synthesis. Early approaches to kinship face synthesis [36], [19], [47] primarily rely on supervised learning for parent-child mapping. ChildPredictor [56] introduces a disentangled representation learning framework to isolate genetic influences, enhancing parental trait transfer. However, the scarcity and low quality of kinship datasets make these methods prone to overfitting, often resulting in low-quality images. Recent works [31], [33], [8], [41] leverage StyleGAN for high-resolution child face generation through latent interpolation. StyleDNA [33] learns parent-to-child mapping in the W space via supervised learning, while KinStyle [8] employs an Image Encoder for precise facial trait capture and fine-tunes with real data to better model kinship relationships. However, KinStyle is limited to generating a single outcome per parent pair and struggles with unnatural skin tones. ChildNet [41] leverages attention and mutation mechanisms for fine-grained control over age, gender, and parental dominance, but its reliance on dropout yields limited diversity. StyleGene [31] introduces a regional facial gene extraction framework with the gene pool to enhance diversity, but lacks fine-grained control over age and gender, exhibiting a trade-off between attribute control and fidelity. Despite these advancements, achieving an optimal balance between diversity and fidelity while enabling continuous age and gender manipulation remains a challenge.

To address these limitations, we propose StyleDiT, a framework that integrates the strengths of StyleGAN and diffusion model. Our approach enables the generation of diverse, high-fidelity kinship faces with fine-grained and continuous control over age and gender. Additionally, our Relation Trait Guidance (RTG) provides independent control

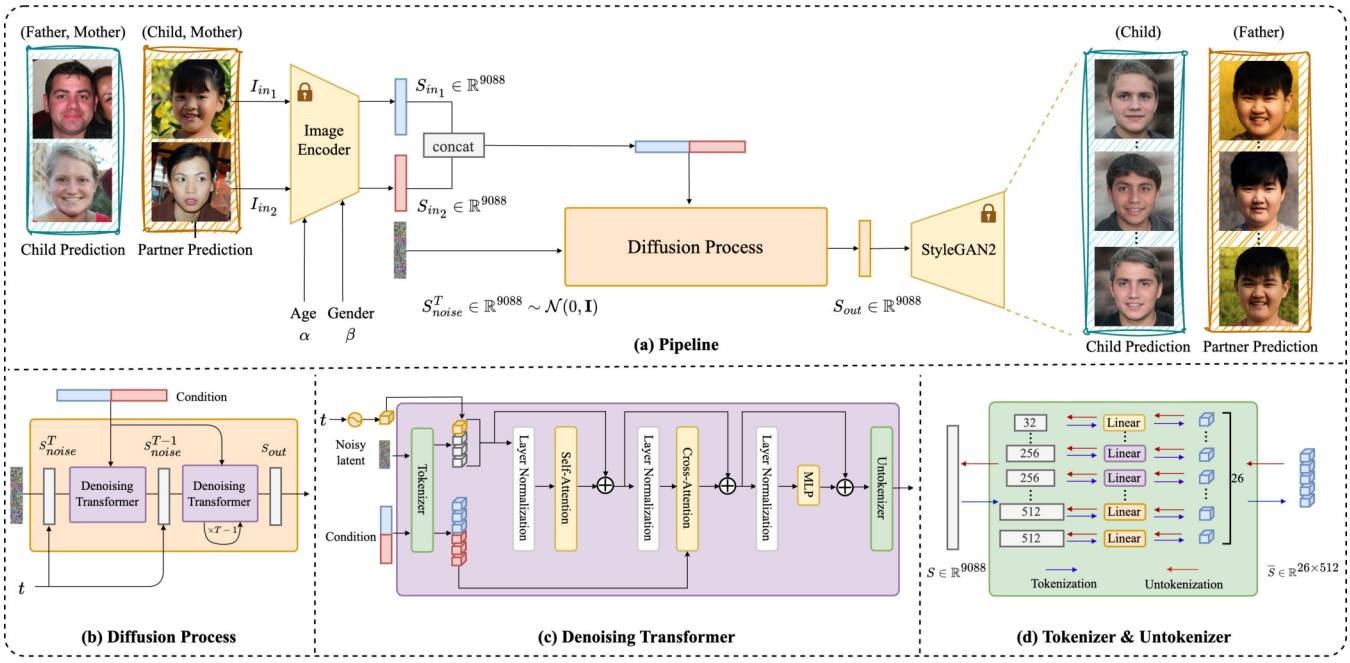


Fig. 3. **Overview of the Proposed Framework.** For both child and partner face prediction tasks, the input images, I_{in_1} and I_{in_2}, are first encoded using an Image Encoder. The encoded style latents, S_{in_1} and S_{in_2}, serve as conditions for the Diffusion Process. A sampled noisy latent is then processed through multiple Denoising Transformer blocks, producing the predicted face latent S_{out}. Finally, StyleGAN2 decodes S_{out} to generate a high-fidelity kinship face. This design effectively reduces the training complexity while ensuring both generation diversity and quality. The lock icon indicates components that remain frozen during training.

over each attribute, allowing users to balance diversity and fidelity according to their specific needs.

III. METHOD

To our knowledge, the proposed method is the first unified framework designed for dual capabilities: synthesizing child faces from parental faces and generating partner faces using a child's face and one parent's image. A notable feature of our method is its ability to produce diverse results while maintaining high fidelity in both tasks and offering fine-grained and continuous attribute control.

A. Preliminary

Child and Partner Face Synthesis. For the task of predicting the child faces, denoted as I_C, our method requires a pair of parental face images: the father's face, I_F, and the mother's face, I_M. In the task of synthesizing the partner faces, represented as $I_{P_{out}}$, the input consists of the child's image, I_C, and a partner face image, $I_{P_{in}}$. In this context, $I_{P_{out}}$ and $I_{P_{in}}$ refer to either the father's image (I_F) or the mother's image (I_M), depending on the target partner face. Our framework is designed to synthesize kinship faces with high fidelity and diversity for both tasks with a specified age α and gender β. Formally, the objective of our framework F, which predicts a face based on two input images I_{in_1} and I_{in_2}, age α and gender β, is defined as follows:

$$I_{out} = F(I_{in_1}, I_{in_2}, \alpha, \beta). \tag{1}$$

Diffusion Model. Diffusion models [48], [22] generate images by iteratively denoising an initial Gaussian noise sample. They consist of a forward process, which progressively adds noise to a clean image, and a backward process, which gradually removes noise to reconstruct the original image. In the forward process, a clean image x_0 is transformed as $x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$ represents Gaussian noise, and $\bar{\alpha}_t$ defines the noise schedule. The backward process refines the noisy sample x_T step by step using a neural network $\epsilon_\theta(x_t, t, c)$, which predicts the added noise while conditioned on c, an auxiliary input that guides generation. The denoising process reconstructs a cleaner image at each step by subtracting the predicted noise. The model is trained by minimizing the Mean Squared Error (MSE) between the predicted and actual noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, \epsilon_t, t, c}\left[|\epsilon_\theta(x_t, t, c) - \epsilon_t|^2\right]. \tag{2}$$

We extend a transformer-based diffusion model [39], [17], modifying it to condition on the one-dimensional style latents of parents as input for realistic child face synthesis, or a child and a parent as input for partner face synthesis.

B. Pipeline

Motivated by previous works [38], [32] that integrates the strengths of StyleGAN and diffusion model, and recognizing the importance of precise control over age and gender in kinship tasks, our proposed framework, illustrated in Fig. 3a, combines the fine-grained and continuous attribute control of StyleGAN's style latent space with the diverse generative capabilities of diffusion models to effectively fit complex distribution. In our framework, input face images I_{in_1} and I_{in_2} are first encoded into style latent codes, S_{in_1} and S_{in_2}, using our pre-trained Image Encoder for precise age and gender control. For further details on the pre-trained Image

Encoder, please refer to Appendix G. Next, we introduce StyleDiT, which conditions a transformer-based diffusion model on these style latents. By leveraging the diffusion model's strength in capturing sophisticated kinship relationship distributions, StyleDiT uses S_{in_1} and S_{in_2} to generate a diverse set of predicted latent codes, S_{out}. This approach overcomes the limited diversity of previous methods [33], [8], [41]. Finally, StyleGAN2 decodes S_{out} to produce the output image, I_{out}.

C. Style Latent Diffusion Transformer

We present Style Latent Diffusion Transformer (StyleDiT), an innovative architecture tailored for kinship tasks that effectively models the style latent space of StyleGAN. In our framework, a style latent is generated through a diffusion process conditioned on input latents, as depicted in Fig. 3b. The transformer architecture, denoted as $\mathcal{T}$, serves as the denoising network. As shown in Fig. 3c, the transformer takes a random Gaussian noise $S^T_{noise} \sim \mathcal{N}(0, \mathbf{I})$, along with conditional inputs S_{in_1} and S_{in_2}, and iteratively denoises S^T_{noise} to $S^0_{noise} = S_{out}$.

At each denoising timestep t, S^t_{noise}, S_{in_1}, and S_{in_2} are processed through a tokenizer (see Fig. 3d). The tokenizer organizes the style latents into distinct groups, each governing specific facial feature. Based on StyleSpace [51], the style latent dimension $S \in \mathbb{R}^{9088}$ represents the complete set of style parameters controlling feature maps and tRGB blocks in StyleGAN2 [27]. Since StyleGAN2 employs different style parameters across layers, these parameters are first divided into 26 groups from the 9,088-dimensional latent space ($\mathbb{R}^{9088}$). Each group is then projected onto tokens of a uniform 512-dimensional embedding space using separate linear layers. Consequently, 26 unique linear layers are used for this transformation, producing tokenized representations $\overline{S} \in \mathbb{R}^{26\times512}$. The sinusoidal embedding of timestep t is projected through an additional linear layer to form a timestep token, which is combined with the noisy token $\overline{S}^t_{noise}$ and summed with a learnable positional encoding vector before being processed by the transformer.

Additionally, the conditional latents $\overline{S}_{in_1}$ and $\overline{S}_{in_2}$ are concatenated, summed with the same learnable positional encoding vector as $\overline{S}^t_{noise}$, and integrated into the denoising process via a cross-attention mechanism. The transformer then outputs denoised tokens $\overline{S}^{t-1}_{noise} \in \mathbb{R}^{26\times512}$, which are converted back into the 9088-dimensional space ($S^{t-1}_{noise} \in \mathbb{R}^{9088}$) using an untokenizer, consisting of another set of linear layers.

During training, we use a kinship dataset consisting of father-mother-child triplets. For child prediction, the training process involves applying noise to the child's initial latent S through a forward diffusion process. A neural network is trained to predict the denoised style latent S^* given the parental latents as conditioning. The model is optimized using MSE loss to minimize the difference between S^* and the original style latent S. Notably, the Image Encoder and StyleGAN2 remain frozen throughout training. During inference, diverse outputs are generated by sampling different instances of S^T_{noise} while keeping S_{in_1} and S_{in_2} constant. This allows StyleDiT to leverage the capability of diffusion models, producing a varied set of realistic kinship faces.

D. Multi-Conditional Classifier-Free Guidance

Our innovative framework, built on a diffusion process, excels at generating diverse outputs from a single pair of input images. This versatility addresses the limited diversity of prior methods [33], [8], [41]. To balance fidelity and diversity, we introduce Relational Trait Guidance (RTG), motivated by Classifier-Free Guidance [23], [6], [4]. RTG provides independent control over each condition, allowing users to manage the trade-off between realism and variability based on their specific requirements.

Relational Trait Guidance (RTG). RTG enables simultaneous use of conditional and unconditional models, blending their predictions during inference. Our conditional diffusion model follows a dual-style latent conditioning approach: $S^*_{\mathcal{T}}(S^t_{noise}|\{S_{in_j}\}^2_{j=1})$, where each condition can be independently adjusted [6], [4]. During training, each condition S_{in_j} is replaced by a null condition $\emptyset$ with a fixed probability. In inference, we compute the guidance direction of each condition as $\Delta^t_j = S^*_{\mathcal{T}}(S^t_{noise}|S_{in_j}) - S^*_{\mathcal{T}}(S^t_{noise}|\emptyset)$. These directions are then weighted and combined using two guidance scales $\{g_j\}^2_{j=1}$:

$$\hat{S}^*_{\mathcal{T}}(S^t_{noise}|\{S_{in_j}\}^2_{j=1}) = S^*_{\mathcal{T}}(S^t_{noise}|\emptyset) + \sum_{j=1}^{2} g_j \Delta^t_j. \quad (3)$$

In the training phase, we randomly assign $S_{in_1} = \emptyset$ for 10% data, $S_{in_2} = \emptyset$ for 10% data, and both $S_{in_1} = \emptyset$ and $S_{in_2} = \emptyset$ for 1% data. The null condition $\emptyset$ symbolizes a learnable style latent corresponding to various age and gender group combinations, derived from our training data. For a detailed design of null conditions, please refer to Appendix C.

IV. EXPERIMENTS

We first outline our experimental settings, including datasets, baselines, and implementation details, in Section IV-A. Section IV-B presents qualitative results, demonstrating that RTG's ability to synthesize diverse, high-fidelity faces. We compare it with baselines for child face prediction and partner face synthesis. Section IV-C provides quantitative evaluations of diversity, kinship verification accuracy, and identity similarity to determine which method best balances diversity and fidelity. In addition, we evaluate the effectiveness of attribute control and conduct a user study on child and partner face prediction. Finally, Section IV-D presents ablation studies on real data training and assesses the impact of the diffusion process and RTG.

A. Experimental Settings

Datasets. We construct a synthetic dataset to alleviate the limitations of real kinship datasets, including issues with resolution, quality, quantity, and diversity. Our approach is motivated by prior studies [8], [31], [41] suggesting that a child's facial features can be inferred through linear interpolation of parental traits. Additionally, medical

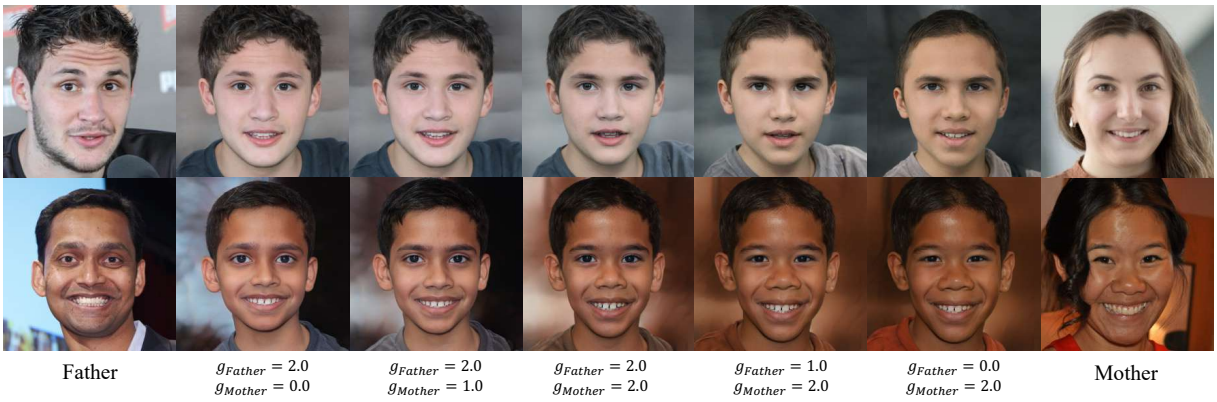


Fig. 4. **The Effect of Different RTG Scales During Inference.** Progressing from left to right, each image results from varying pairs of guidance scales, with higher scales producing outcomes that more closely resemble the specified conditions.

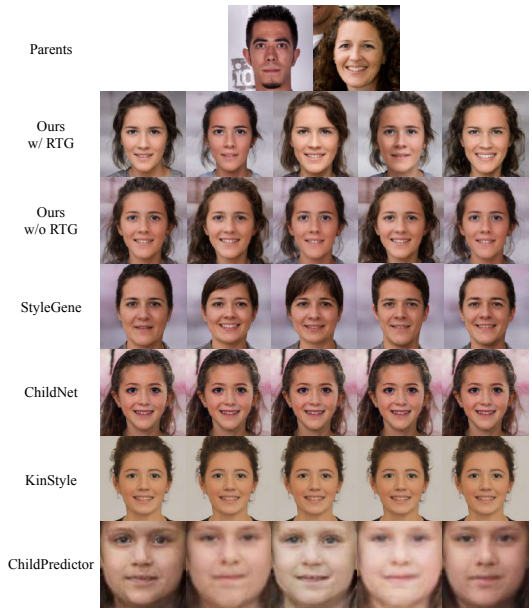


Fig. 5. **Demonstrating RTG's Impact on Facial Diversity.** The figure showcases the effectiveness of RTG in our method by comparing the diversity of synthesized results from our approach, StyleGene [31], ChildNet [41], KinStyle [8], and ChildPredictor [56]. All faces are generated with a fixed age of 15 years and a female attribute.

research [12], [2] supports the notion that offspring inherit genes through random genetic combinations, reinforcing the feasibility of interpolating parental traits to generate child faces. Our dataset consists of 100,000 synthetic father-mother-child triplets, generated as follows: We first sample 70,000 male-female pairs from CelebA-HQ as parental pairs, map them into StyleGAN2's latent space, and apply 200 attribute combinations (age 0-99 × two genders). Child faces are synthesized via linear interpolation of parental latents, with varying resemblance weights to enhance diversity. Additionally, 30,000 male-female pairs are randomly sampled from StyleGAN2's latent space and processed similarly. This dataset is used for training both child and partner face prediction. Visual examples are offered in Appendix Fig. S.7.

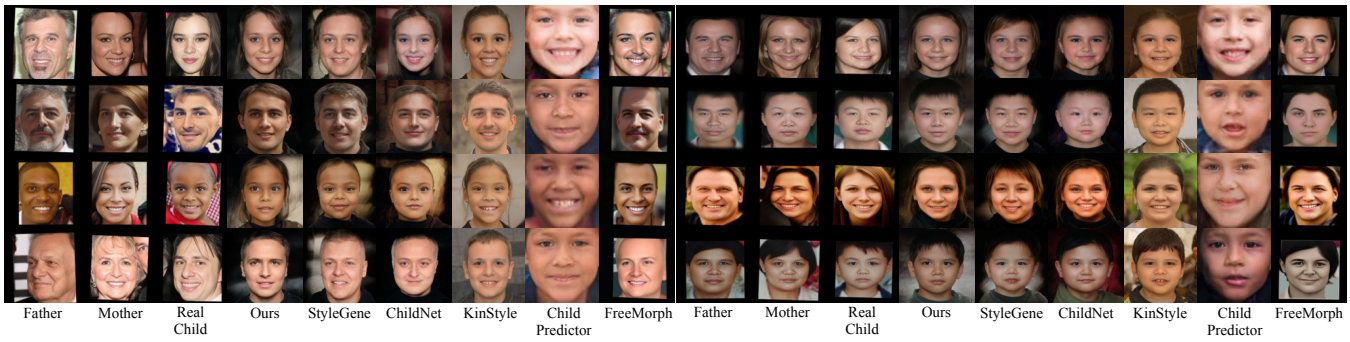
For evaluation, we use the test splits of FIW [44], TSKinFace [43], and FF-Database [56]. We preprocess the images with facial alignment [14], enhancement [57], and resizing to 256×256 pixels. Since FIW and TSKinFace lack predefined test sets and FF-Database provides only training/validation splits, we define 296, 149, and 368 test triplets, respectively.

Baselines. We compare our method with state-of-the-art child face synthesis models, including StyleGene [31], ChildNet [41], KinStyle [8], and ChildPredictor [56], replicating the results using their official implementations. Moreover, we also compare with the recent image morphing method FreeMorph [7]. For partner face synthesis, due to public unavailability of the official source code, we mainly qualitatively compare the generated faces available from the paper of ParentGAN [16] with our approach.

Implementation Details. We train our model using the AdamW optimizer with a batch size of 1,000, a learning rate of 0.001, and 4,000 epochs with 500 diffusion timesteps. Training is conducted on 4 NVIDIA RTX A5000 GPUs and completes in approximately 20 hours. During evaluation, all synthesized kinship faces are generated using a DDIM sampler [49] with 50 steps. For child prediction, the guidance scale is set to 1.2 for parents, while for partner prediction, it is set to 1.2 for the child and 0.0 for the parent.

B. Qualitative Evaluation

Effectiveness of Relational Trait Guidance. Our proposed Relational Trait Guidance (RTG) enables a flexible trade-off between diversity and fidelity in generated images. As illustrated in Fig. 5, our method with RTG produces diverse child predictions, whereas disabling RTG reduces variation. StyleGene[31] generates diverse faces but lacks precise control over facial attributes, often leading to inaccurate age and gender when maintaining parental resemblance. ChildNet [41] achieves high-fidelity child face generation with accurate age and gender but offers limited diversity. KinStyle [8] produces high-fidelity kinship faces but follows a deterministic approach, generating only a single prediction per input and frequently exhibiting unnatural skin tones. ChildPredictor [56] generates diverse child faces but with lower image quality than other approaches. Furthermore, as shown in Fig. 4, RTG enables fine-grained resemblance



(a) FIW dataset (b) TSKinFace dataset

Fig. 6. **Qualitative Comparison of Synthesized Child Faces.** The first three columns in both (a) FIW and (b) TSKinFace datasets show the father, mother, and real child, followed by synthesized child images from six different baselines. The results include faces across different races and skin tones, demonstrating the framework's ability to generalize across diverse racial backgrounds. Each method generates child faces based on the real child's gender and age within each family.

control, allowing child faces to be selectively guided toward either the father or the mother by adjusting their respective guidance scales during inference.

Comparison with the State-of-the-Art. Fig. 6 presents qualitative results for child prediction on the FIW and TSKinFace datasets, comparing our approach with StyleGene [31], ChildNet [41], KinStyle [8], ChildPredictor [56], and FreeMorph [7]. StyleGene generates diverse, high-quality child faces but struggles to balance fidelity to parental features with accurate age and gender representation within its gene pool framework. ChildNet generates high-quality faces but offers limited diversity. KinStyle produces high-fidelity faces with well-preserved parental identity, but its deterministic framework limits diversity and often results in inaccurate skin tones. ChildPredictor struggles to generate high-quality images due to the absence of a StyleGAN-based generator. FreeMorph often introduces incoherent artifacts and does not provide explicit age and gender control, which limits its suitability for kinship synthesis. In contrast, our method achieves an optimal balance, generating diverse, high-quality images that maintain strong parental traits across various racial backgrounds while providing fine-grained control over age and gender.

Partner Face Synthesis. Beyond child face synthesis, our framework also facilitates high-fidelity partner face generation. Using the same datasets and model architecture, it synthesizes a partner's face given a child image and the other partner's image. Fig. 7 qualitatively compares visual fidelity on the left and illustrates the diversity of the generated partner faces on the right. From the results, we observe that our proposed method generates clearer facial images than ParentGAN [16]. Moreover, when given Asian faces as input, ParentGAN often generates Western-looking faces and fails to preserve racial characteristics, whereas our approach consistently maintains them.

C. Quantitative Evaluation

We evaluate our approach using the test splits of the FIW, TSKinFace, and FF-Database datasets. For each family, we generate 20 children for child prediction and 20 partners for partner prediction. The generated images are assessed using a

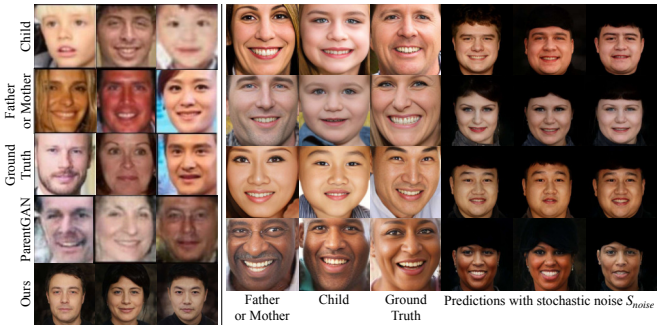


Fig. 7. **Qualitative Results of Partner Face Synthesis.** A comparison with ParentGAN [16] is shown on the left, illustrating the superior visual quality and fewer artifacts in our synthesized partner faces, while the right side demonstrates the diversity of generated partners across different races.

comprehensive evaluation framework that includes diversity, kinship verification accuracy, and identity similarity. Results for child prediction are presented below, while partner prediction results are provided in Appendix B.2.

Diversity Measurement. We evaluate the diversity of synthetic children using LPIPS [55] and the mean Diversity Score (mean DS). LPIPS computes the L1 distance between image feature pairs extracted by AlexNet [30], pre-trained on ImageNet [13]. To obtain the LPIPS score, we first calculate feature distances among the 20 synthetic faces within each family and then average these values across all families. Higher LPIPS scores indicate greater diversity. The mean DS measures the pairwise cosine similarity of extracted face representations within each family, averaged across all families. To capture critical semantic and facial details, we use ArcFace [15], AdaFace [28], and TransFace [11], pre-trained on MSCeleb1M [21], WebFace12M [58], and Glint360K [3], respectively. The DS metric is defined as follows:

$$DS = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} \tag{4}$$

where N is the number of samples and x_i, x_j are individual feature vectors. A lower mean DS signifies greater diversity, indicating that generated images exhibit less resemblance to each other despite sharing the same input conditions.

Table I compares LPIPS and DS scores for StyleGene,

TABLE I

QUANTITATIVE EVALUATION OF SYNTHESIZED CHILD FACES. WE REPORT LPIPS, DIVERSITY (MEAN DS), KINSHIP VERIFICATION ACCURACY (ACC), AND IDENTITY SIMILARITY (ID SIM). OUR METHOD ACHIEVES A STRONG BALANCE BETWEEN DIVERSE SYNTHESIS AND HIGH FIDELITY. MEAN DS AND ID SIM VALUES ARE REPORTED IN THE ORDER CORRESPONDING TO ARCFACE, ADAFACE, AND TRANSFACE REPRESENTATIONS.

Methods	**LPIPS (↑)**			**mean DS (↓)**		
	FIW	TSKinFace	FF-Database	FIW	TSKinFace	FF-Database
FreeMorph [7]	–	–	–	–	–	–
ChildPredictor [56]	0.1953	0.1944	0.1868	0.7805/0.6651/0.7498	0.7811/0.6671/0.7435	0.7702/0.6645/0.7460
KinStyle [8]	–	–	–	–	–	–
ChildNet [41]	–	–	–	–	–	–
StyleGene [31]	0.1375	0.1219	0.1347	0.6757/0.6054/0.6106	0.7496/0.6869/0.6913	**0.6966/0.6282/0.6407**
StyleDiT (Ours)	**0.2030**	**0.2206**	**0.2059**	**0.6140/0.5638/0.5708**	**0.6780/0.6237/0.6268**	0.7078/0.6436/0.6691
Methods	**ACC (↑)**			**ID Sim (↑)**		
	FIW	TSKinFace	FF-Database	FIW	TSKinFace	FF-Database
FreeMorph [7]	53.61	51.26	61.58	0.6141/0.6004/0.5803	0.6233/0.6140/0.6291	0.6384/0.6378/0.6254
ChildPredictor [56]	50.00	50.00	49.99	0.4996/0.5037/0.4964	0.4989/0.4977/0.4961	0.4979/0.4951/0.4983
KinStyle [8]	**68.10**	**67.79**	**77.24**	**0.7707/0.7579/0.7460**	**0.7768/0.7990/0.7787**	**0.7772/0.7423/0.7426**
ChildNet [41]	64.63	63.10	61.47	0.6783/0.6815/0.6496	0.7041/0.7284/0.7457	0.6684/0.6611/0.6568
StyleGene [31]	64.61	62.57	65.40	0.7132/0.6937/0.6840	0.7242/0.7504/0.7448	0.6887/0.6739/0.6903
StyleDiT (Ours)	64.64	65.09	69.90	0.7003/0.6970/0.6936	0.7244/0.7539/0.7487	0.7124/0.7154/0.7114

ChildPredictor, and our method. KinStyle is excluded due to its single-result limitation, and ChildNet and FreeMorph are omitted as they produce minimal diversity. Our method achieves the highest LPIPS across all three datasets and the lowest DS in FIW and TSKinFace, with the second-lowest in FF-Database, demonstrating its superior ability to generate diverse child face images compared to other baselines.

Kinship Verification. We use kinship verification accuracy to assess whether a genetic relationship exists between synthetic children and their parents. Higher accuracy indicates that the generated descendants appear more realistic. To this end, we employ a kinship classifier [40], which utilizes ResNet101 with ArcFace as the backbone. Specifically, we randomly synthesize 20 faces per family. Positive pairs consist of images of the same person, while negative pairs consist of images of different individuals, and accuracy is computed accordingly. As shown in Table I, our method achieves a strong balance between diversity and kinship verification accuracy. Although KinStyle attains higher accuracy, its deterministic design restricts it to producing only one output per input.

Identity Similarity. To evaluate the fidelity of our kinship face synthesis methods, we use the Area Under the Receiver Operating Characteristic (AUC-ROC) curve as the primary metric for measuring identity similarity between predicted and ground truth children. This curve measures the model's ability to synthesize high-fidelity child and partner faces by plotting the true positive rate against the false positive rate at various discrimination thresholds. For a more robust assessment, we extract facial features using the same models employed for computing mean DS, ensuring a comprehensive evaluation of the generated faces.

We randomly generate 20 faces per family, forming positive and negative pairs based on identity, and compute the identity similarity accordingly. Table I shows that our method achieves a balanced trade-off between diversity and identity similarity. While KinStyle attains higher identity similarity on child prediction, its deterministic framework restricts its ability to produce diverse outputs.

TABLE II

PERFORMANCE COMPARISON OF ATTRIBUTE CONTROL. STYLEGENE † REFERS TO A 10% UTILIZATION OF THE GENE POOL IN THE LINEAR COMBINATION OF PARENTS AND GENE POOL, WHILE STYLEGENE ‡ DENOTES 90% USAGE. STYLEDIT APPLIES A PARENTAL GUIDANCE SCALE OF 2.0 FOR BOTH PARENTS. OUR METHOD ACHIEVES A MORE BALANCED PERFORMANCE BETWEEN ATTRIBUTE CONTROLLABILITY AND KINSHIP RESEMBLANCE, OUTPERFORMING STYLEGENE IN IDENTITY SIMILARITY WHILE MAINTAINING SUPERIOR AGE AND GENDER ACCURACY.

	Age MSE (↓)	Gender Acc. (↑)	ACC (↑)	ID Sim (↑)
StyleDiT (Ours)	**0.0034**	**99.10**	72.35	**0.6996**
StyleGene †	0.0164	87.99	**74.48**	0.6771
StyleGene ‡	0.0045	97.93	58.04	0.6685

Effectiveness of Fine-grained Age and Gender Control. Our framework leverages StyleGAN's latent space to enable fine-grained attribute control. Similarly, StyleGene generates kinship faces with attribute control; however, its approach relies on a diverse gene pool spanning multiple ages, genders, and races. This reliance can compromise input fidelity, especially when extensive gene pool information is required for precise attribute manipulation.

To evaluate our framework's ability to achieve both precise attribute control and high fidelity, we conduct a comparative analysis using the test split of the FF-Database dataset. For each family, we generate 20 child images. As shown in Table II, we assess age control using MSE and gender control using binary classification accuracy. Fidelity is measured through kinship verification accuracy (ACC) and identity similarity (ID Sim). The results demonstrate that our method excels in both attribute precision and fidelity. Moreover, they highlight an inherent trade-off in StyleGene, where increasing the gene pool's usage improves attribute control but reduces fidelity. For visual examples illustrating precise age and gender control, see Appendix Fig. S.5.

User Study. To further demonstrate the effectiveness of our method, we conduct a user study with 73 participants. Each evaluator is presented with child faces generated by

TABLE III

USER STUDY RESULTS FOR CHILD PREDICTION. THE TABLE PRESENTS THE AVERAGE RANKING OF DIFFERENT APPROACHES IN THE CHILD PREDICTION TASK. OUR APPROACH OUTPERFORMS EXISTING BASELINES, DEMONSTRATING SUPERIOR PERCEPTUAL QUALITY AND RESEMBLANCE TO BOTH PARENTAL AND CHILD TRAITS.

	StyleDiT (Ours)	StyleGene	ChildNet	KinStyle	ChildPredictor
Child Avg. Rank (↓)	**2.42**	2.44	3.05	2.75	4.44

TABLE IV

PERFORMANCE COMPARISON OF USING REAL DATA. WE EVALUATE IDENTITY SIMILARITY (ID SIM) ACROSS FOUR CONFIGURATIONS: *(i)* OUR DEFAULT SYNTHETIC-ONLY SETUP, *(ii)* SYNTHETIC + REAL TRAINING, *(iii)* REAL-DATA FINE-TUNING, AND *(iv)* REAL-ONLY TRAINING. OUR DEFAULT SETTING CONSISTENTLY ACHIEVES HIGHER IDENTITY SIMILARITY THAN THE REAL-DATA VARIANTS. VALUES ARE REPORTED IN THE ORDER CORRESPONDING TO FACE REPRESENTATIONS EXTRACTED WITH ARCFACE, ADAFACE, AND TRANSFACE.

Methods	FIW	TSKinFace	FF-Database
StyleDiT (Ours) *(i)*	**0.7003/0.6970/0.6936**	**0.7244/0.7539/0.7487**	**0.7124/0.7154/0.7114**
(ii)	0.6923/0.6821/0.6843	0.6261/0.6609/0.6387	0.6432/0.6336/0.6452
(iii)	0.6817/0.6713/0.6806	0.6198/0.6428/0.6385	0.6398/0.6246/0.6603
(iv)	0.5362/0.5183/0.5143	0.5181/0.5142/0.5112	0.5261/0.5279/0.5222

StyleDiT, StyleGene [31], ChildNet [41], KinStyle [8], and ChildPredictor [56] based on seven parent pairs and ranks them by similarity to the parents. Additionally, a separate assessment evaluates the resemblance between generated and real children, where three children are synthesized from three other parent pairs, and participants rank them by identity similarity. Beyond similarity, participants also assess the generated faces in each question based on quality, realism, and skin color naturalness. As shown in Table III, our method achieves the lowest average ranking, demonstrating its superiority in child face synthesis compared to other baselines. For results on partner face synthesis and more details, please refer to Appendix H.

D. Ablation Study

Effectiveness of Using Real Data. To evaluate the effectiveness of incorporating real data into our framework, we conduct three experiments in addition to our default setting: *(i)* our default synthetic-only setup, *(ii)* training with a combination of synthetic and real kinship data, *(iii)* fine-tuning on real kinship data, and *(iv)* training exclusively on real kinship data. However, these settings did not yield overall performance improvements. As shown in Table IV, these approaches resulted in lower identity similarity compared with our default setup. We hypothesize that the limited availability and low quality of existing real-world kinship datasets constrain the model's capacity to effectively learn complex kinship distributions. Even applying identity-preserving super-resolution [57] does not appear to alleviate this limitation. In addition, qualitative comparisons in Fig. 8 demonstrate that real-data-based training *(ii)*-*(iv)* sometimes capture coarse kinship-related facial structures, but our default setting consistently produces more stable and visually convincing results. Consequently, all reported results in the main paper are based on our default setting, which we identify as a better setting. For Table IV and Fig. 8, we employ the FF-Database containing 8,187 kinship pairs as the real dataset. The limitations arising from the insufficient quantity and quality of existing kinship datasets will be further explored in future work. More experimental details of real data utilization are provided in Appendix E.

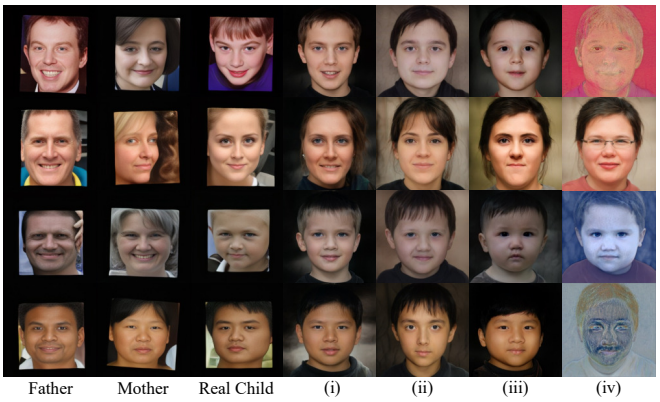


Fig. 8. **Qualitative Comparison of Using Real Data.** We compare three real-data settings against our default: *(i)* our default setup, *(ii)* training on a combination of synthetic and real kinship data, *(iii)* fine-tuning on real kinship data, and *(iv)* training solely on real kinship data. The default consistently produces more stable and visually convincing results, whereas the others may capture coarse facial contours in some cases.

TABLE V

EFFECTIVENESS OF DIFFUSION PROCESS AND RTG. WE ASSESS THE IMPACT OF THE DIFFUSION PROCESS AND RTG. THE DIFFUSION PROCESS PROVIDES A SLIGHT IMPROVEMENT IN DIVERSITY, WHILE RTG NOTABLY ENHANCES DIVERSITY IN SYNTHESIZED CHILDREN AND PRESERVES COMPETITIVE IDENTITY SIMILARITY. VALUES ARE REPORTED IN THE ORDER CORRESPONDING TO FACE REPRESENTATIONS EXTRACTED WITH ARCFACE, ADAFACE, AND TRANSFACE.

Transformer	Diffusion	RTG	ID Sim (↑)	mean DS (↓)
✓	✓	✓	0.7003/0.6970/0.6936	**0.6140/0.5638/0.5708**
✓	✓		0.7636/0.7469/0.7373	0.9984/0.9978/0.9980
✓			0.7666/0.7428/0.7360	—

Effectiveness of Diffusion Process and RTG. The assessment considers identity similarity (ID Sim) and diversity (mean DS), measured on the test split of the FIW dataset. Table V shows that integrating the diffusion process improves mean DS for synthesized descendants. As expected, incorporating RTG effectively enhances the diversity of generated faces while maintaining a favorable ID Sim.

V. CONCLUSION

We introduce StyleDiT, a unified framework for generating diverse, high-fidelity kinship faces with precise age and gender control. Our Relational Trait Guidance (RTG) enables independent control of influencing factors, allowing users to balance diversity and fidelity effectively. Extensive benchmark evaluations and user studies show that StyleDiT outperforms state-of-the-art methods, achieving an optimal trade-off between diversity and fidelity, underscoring its effectiveness and innovation.

REFERENCES

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] A. Alvergne, C. Faurie, and M. Raymond. Differential facial resemblance of young children to their parents: who do children look like more? *Evolution and Human behavior*, 28(2):135–144, 2007.

[3] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[4] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, 2023.

[6] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 18392–18402, 2023.

[7] Y. Cao, C. Si, J. Wang, and Z. Liu. Freemorph: Tuning-free generalized image morphing with diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[8] L.-C. Cheng, S.-C. Hsu, P.-H. Lee, H.-C. Lee, C.-H. Lin, J.-C. Chen, and C.-Y. Wang. Kinstyle: A strong baseline photorealistic kinship face synthesis with an optimized stylegan encoder. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022.

[9] J. B. Cole, M. Manyama, J. R. Larson, D. K. Liberton, T. M. Ferrara, S. L. Riccardi, M. Li, W. Mio, O. D. Klein, S. A. Santorico, et al. Human facial shape and size heritability and genetic correlations. *Genetics*, 205(2):967–978, 2017.

[10] Y. Dalva, H. Yesiltepe, and P. Yanardag. Gantastic: Gan-based transfer of interpretable directions for disentangled image editing in text-to-image diffusion models. *arXiv preprint arXiv:2403.19645*, 2024.

[11] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20642–20653, 2023.

[12] L. M. DeBruine, B. C. Jones, A. C. Little, and D. I. Perrett. Social perception of facial resemblance in humans. *Archives of sexual behavior*, 37:64–77, 2008.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. Ieee, 2009.

[14] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5203–5212, 2020.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] M. M. Emara, M. Farouk, and M. W. Fakhr. Parent gan: image generation model for creating parent's images using children's images. *Multimedia Tools and Applications*, 84(24):28643–28665, 2025.

[17] Z. Erkoç, F. Ma, Q. Shan, M. Nießner, and A. Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[18] R. Gandikota, J. Materzyńska, T. Zhou, A. Torralba, and D. Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *Proceedings of the Eurpean Conference on Computer Vision (ECCV)*, 2024.

[19] P. Gao, S. Xia, J. Robinson, J. Zhang, C. Xia, M. Shao, and Y. Fu. What will your child look like? dna-net: Age and gender aware kin face synthesizer. *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 2021.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.

[21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[23] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[24] O. Kafri, O. Patashnik, Y. Alaluf, and D. Cohen-Or. Stylefusion: A generative model for disentangling spatial segments. *arXiv preprint arXiv:2107.07437*, 2021.

[25] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[26] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[29] J. N. Kolf, T. Rieber, J. Elliesen, F. Boutros, A. Kuijper, and N. Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 806–816, 2023.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[31] H. Li, X. Hou, Z. Huang, and L. Shen. Stylegene: Crossover and mutation of region-level facial genes for kinship face synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[32] X. Li, X. Hou, and C. C. Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2187–2196, 2024.

[33] C.-H. Lin, H.-C. Chen, L.-C. Cheng, S.-C. Hsu, J.-C. Chen, and C.-Y. Wang. Styledna: A high-fidelity age and gender aware kinship face synthesizer. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2021.

[34] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3086–3095, 2023.

[35] K. Narayan, V. VS, R. Chellappa, and V. M. Patel. Facexformer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024.

[36] S. Ozkan and A. Ozkan. Kinshipgan: Synthesizing of kinship faces from family photos by regularizing a deep face network. In *Proceedings of the IEEE international conference on image processing (ICIP)*, 2018.

[37] F. Paraperas Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[38] R. Parihar, S. VS, S. Mani, T. Karmali, and R. V. Babu. Precisecontrol : Enhancing text-to-image diffusion models with fine-grained attribute control. In *Proceedings of the Eurpean Conference on Computer Vision (ECCV)*, 2024.

[39] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[40] J. L. Peng, K. W. Chang, and S.-H. Lai. Kfc: Kinship verification with fair contrastive loss and multi-task learning. In *The British Machine Vision Conference (BMVC)*, 2023.

[41] M. Pernuš, M. Bhatnagar, B. Samad, D. Singh, P. Peer, S. Dobrišek, et al. Childnet: Structural kinship face synthesis model with appearance control mechanisms. *IEEE Access*, 11:49971–49991, 2023.

[42] J. N. Pinkney and C. Li. clip2latent. *arXiv preprint arXiv:2210.02347*, 2022.

[43] X. Qin, X. Tan, and S. Chen. Tri-subject kinship verification: Understanding the core of a family. *IEEE Transactions on Multimedia*,

2015.
[44] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu. Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *Proceedings of the ACM on Multimedia Conference (MM)*, 2016.
[45] H. O. Shahreza and S. Marcel. Hyperface: Generating synthetic face recognition datasets by exploring face embedding hypersphere. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
[46] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
[47] R. Sinha, M. Vatsa, and R. Singh. Familygan: Generating kin face images using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, pages 297–311, 2020.
[48] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.
[49] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
[50] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.
[51] Z. Wu, D. Lischinski, and E. Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[52] X. Yao, A. Newson, Y. Gousseau, and P. Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13789–13798, 2021.
[53] J. Yu, M. Li, X. Hao, and G. Xie. Deep fusion siamese network for automatic kinship verification. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 892–899, 2020.
[54] C. Zhang, Y. Chen, Y. Fu, Z. Zhou, G. Yu, B. Wang, B. Fu, T. Chen, G. Lin, and C. Shen. Styleavatar3d: Leveraging image-text diffusion models for high-fidelity 3d avatar generation. *arXiv preprint arXiv:2305.19012*, 2023.
[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 586–595, 2018.
[56] Y. Zhao, L.-M. Po, X. Wang, Q. Yan, W. Shen, Y. Zhang, W. Liu, C.-K. Wong, C.-S. Pang, W. Ou, et al. Childpredictor: A child face prediction framework with disentangled learning. *IEEE Transactions on Multimedia*, 2022.
[57] S. Zhou, K. Chan, C. Li, and C. C. Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.
[58] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.