Introducing the EG3D Generative Model to Build a High-Fidelity Dynamic Face Actuation System for Gamified Film and Television Content

Wenjun Bao
Digital Entertainment College, Shanghai Film Art Academy, Shanghai 200120, China
Corresponding author: Wenjun Bao (e-mail: cocohelenbao@163.com)

ABSTRACT With the increasing demand for immersion and interactivity in gamified film and television content, maintaining high-fidelity dynamic facial expressions remains challenging due to insufficient detail preservation and temporal inconsistencies across consecutive frames. To address these issues, this paper proposes EG3D-TA, a framework that integrates EG3D implicit field modeling with a Temporal Awareness (TA) actuation network. The proposed method constructs a three-dimensional generative representation based on EG3D, where Neural Radiance Fields (NeRF) implicitly model high-resolution facial geometry, while a keypoint-driven mapping network transforms two-dimensional facial expressions into continuous W+ latent encodings. An optical flow-guided latent interpolation strategy based on PWC-Net is further introduced to optimize inter-frame trajectories and improve temporal coherence. In addition, model pruning and feature caching mechanisms enable low-latency real-time inference. Experimental results demonstrate that the proposed approach achieves high visual fidelity (LPIPS: 0.12–0.19; PSNR: 36.0–38.2 dB) and excellent motion consistency (TVD: 0.08–0.12; FVD: 78.3–88.9), while maintaining an interactive performance of 21.6 FPS at 512^2 resolution with an end-to-end latency of 46.3 ms. Beyond high-fidelity digital character generation, the proposed geometry-aware and temporally consistent representation framework provides valuable methodological support for three-dimensional signal reconstruction, immersive visual communication, and intelligent sensing systems in advanced electromagnetic information processing and future integrated communication–perception applications.

INDEX TERMS implicit field modeling, temporal awareness, geometry-aware 3d generation, electromagnetic information processing, intelligent sensing systems

I. INTRODUCTION

THE evolution of virtual faces is reshaping the emotional dimension of human-computer interaction, a shift toward enhanced digital fidelity. In film and television special effects, virtual anchors, and extended reality scenarios, audiences' expectations of digital characters have moved beyond static realism to the perception of the vitality of expressions [1,2]. A hesitant blink, a subtle twitch of the eye during a smile—these nonverbal signals carry key information about emotions [3]. High-fidelity dynamic faces represent not only a technological milestone but also a crucial means of fostering trust and empathy [4-6]. When a generative system accurately reproduces the rhythm of muscle movements and the continuity of skin deformation, the resulting virtual avatar exhibits behavioral realism that supports human-like perception [7,8].

Some methods attempt to alleviate the jump problem through post-processing filtering, but such operations often sacrifice dynamic response speed, affecting the real-time interactive experience [9]. The computational load of the model during inference also limits the frame rate stability of high-resolution output, especially in mobile or lightweight deployment scenarios [10,11]. These issues combine to cause the generated images to gradually deviate from the boundaries of realism over extended playback, weakening the user's emotional engagement with the virtual character.

This paper develops EG3D-TA, a 3D-aware generation framework for highly dynamic expression-driven face generation, aiming to comprehensively improve the realism and temporal stability of generated sequences without sacrificing visual quality. The core contributions are reflected in three aspects: First, based on the EG3D architecture, a foundation for 3D perception generation is established, and NeRF implicit fields are used to accurately model the high-resolution geometry and texture distribution of the face, ensuring that the single-frame output is rich in

details. Second, a key point-driven latent space mapping mechanism is proposed to accurately convert 2D expression action sequences into continuous encodings in W+ space. By introducing an optical flow-guided latent code interpolation module and using the motion flow extracted by PWC-Net (Pyramid, Warping, Cost Volume–Network), the latent path is dynamically corrected, so that the deformation process between adjacent frames conforms to the laws of real muscle movement. Third, a lightweight strategy for the inference phase was designed, combining structured model pruning with a tri-plane feature caching mechanism to significantly reduce computational redundancy, enabling low-latency, real-time generation while maintaining high output resolution. This approach overcomes the trade-off between dynamic consistency and responsiveness faced by existing generative models, enabling closed-loop control of latent space trajectories using optical flow priors. Experiments demonstrate that this framework outperforms current mainstream solutions in terms of naturalness of expression transitions, detail preservation, and temporal stability, providing a viable technical path for virtual characters.

II. EG3D-TA SYSTEM CONSTRUCTION AND IMPLEMENTATION

The EG3D-TA system architecture in Figure 1 illustrates the complete process of high-fidelity dynamic face actuation. The system takes a 2D facial landmark sequence as input, which is converted into a W+ latent code via a temporal perceptual mapping network. This code drives the EG3D generative backbone, which generates high-resolution facial frames using three-planar features and neural radiance field volume rendering. The core innovation lies in the optical flow-guided optimization module, which uses optical flow extracted by PWC-Net to calculate consistency loss and reversely optimize latent code interpolation, forming a closed-loop feedback loop that improves temporal consistency. Simultaneously, model pruning, quantization, and feature caching strategies significantly accelerate inference, ensuring the system achieves real-time interactivity while maintaining cinematic quality.

IMPLICIT 3D FACE REPRESENTATION MODELING BASED ON EG3D

3D Perception Generative Modeling Based on the EG3D Architecture

To achieve high-fidelity dynamic face generation, the system uses the EG3D network as the generative backbone to construct an implicit representation model with 3D perception capabilities. Compared to pure implicit dynamic NeRF or K-Planes, EG3D's three-plane representation maintains NeRF-level geometric fidelity while providing structured, differentiable, and cacheable intermediate features, which lays a critical foundation for subsequent time-aware driving and real-time optimization. This architecture employs three-plane features as an intermediate geometric representation, mapping the latent vector to three orthogonal planes, corresponding to the front-back, left-right, and top-bottom spatial distribution of the face. The feature maps on each plane are sampled using bilinear interpolation and used to drive the subsequent NeRF network for volume rendering. During training, a dataset of facial images captured from multiple perspectives, covering different poses, lighting conditions, and expressions, is input to ensure that the model learns a three-dimensional prior that decouples pose and expression. The network embeds the target pose parameters (represented by pitch, yaw, and roll angles) and the expression latent variables into the mapping network through conditional input, guiding the generator to output a high-resolution image from the corresponding perspective. The NeRF module discretizes query points in space along the camera's line of sight using ray sampling. Each point is predicted to have its density and RGB (Red, Green, Blue) color value through the MLP (MultiLayer Perceptron) [12,13] network, and finally synthesizes a 2D image using the volume rendering integral formula:

$$C(r) = \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) c_i \tag{1}$$

r represents the camera ray, T_i is the cumulative transmittance, σ_i is the density of the i-th sampling point, δ_i is the distance between adjacent sampling points, and c_i is the corresponding color output. This rendering mechanism ensures that the generated images are viewpoint-consistent while preserving microscopic details such as skin texture, pores, and fine lines. During the training phase, an adversarial loss and a perceptual loss are jointly optimized. The discriminator is supervised in both image space and 3D viewpoint consistency, forcing the generated results to maintain geometric stability and texture continuity across different poses.

LATENT SPACE DECOUPLING AND MULTI-CONDITION DRIVEN GENERATION MECHANISM

To further improve generation quality and control accuracy, the model introduces an explicit decoupling mechanism for pose and expression in the latent space. By separating the W+ spatial encoding output by the mapping network, the overall latent code is decomposed into pose-related and expression-related components, which are respectively fed into the three-plane feature generation module and the NeRF color branch. The pose component governs global head pose and camera viewpoint, whereas the expression component modulates local geometric offsets and corresponding texture variations driven by facial muscle activity. During the inference phase, the system receives external input of pose angle sequences and expression driving signals. Using a pre-trained pose-expression decoupling module, it maps these into corresponding latent space offsets, enabling fine-grained control over generated content. To enhance the geometric accuracy of the generated results, the model introduces a depth map supervision signal during training,

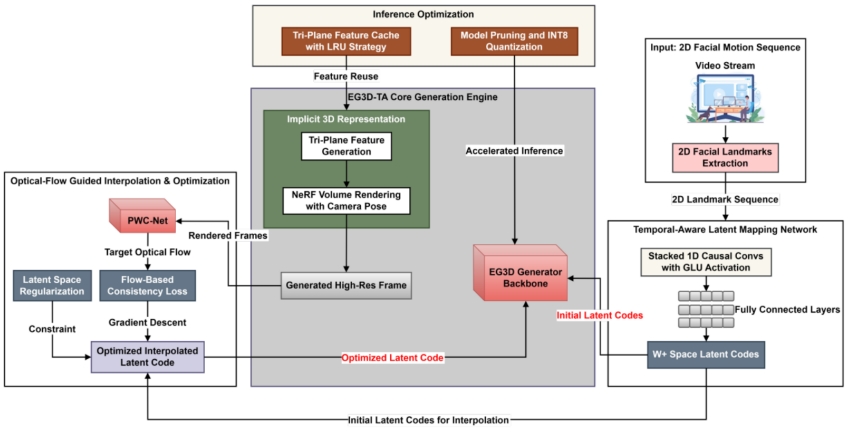

FIGURE 1. EG3D-TA System Architecture

uses a monocular depth estimation algorithm to generate an approximate depth map of the real face, and imposes an L1 distance constraint on the depth output predicted by NeRF. The depth consistency loss can be expressed as:

$$L_{\text{depth}} = \mathbb{E}_{x\sim D}\left[\|D_{\text{pred}}(x) - D_{\text{gt}}(x)\|_1\right] \tag{2}$$

Dpred is the depth map rendered using NeRF, Dgt is the estimated true depth, and D is the training data distribution. This constraint effectively mitigates the problem of structural collapse caused by occlusion and non-rigid deformation in areas such as the cheekbones and eye sockets during large-angle sideways movements or exaggerated expressions. Furthermore, the spatial orthogonality of the three-plane features ensures the stability of feature sampling across different viewpoints, avoiding the geometric distortion that occurs in traditional NeRF at sparse viewpoints [14,15]. Ultimately, this 3D perceptual generative model can simultaneously output high-resolution images and corresponding depth maps without explicit mesh modeling, providing a precise joint geometry-texture representation foundation for subsequent dynamic driving.

DESIGN OF A LATENT SPACE MAPPING NETWORK DRIVEN BY FACIAL KEYPOINTS

End-to-end Mapping Architecture from Keypoint Sequence to Latent Space

Each convolutional layer is followed by batch normalization and gated linear units (GLUs) to enhance nonlinear representation while suppressing redundant responses. After multiple layers of downsampling, the features output a compact temporal aggregation vector, which is then projected through a fully connected layer into a latent code in W+ space corresponding to the current frame. This latent code is directly injected into the generative backbone of EG3D as the core input to control facial geometry and texture variations. To ensure the physical plausibility of the mapping process, expression parameters from the FLAME (Faces Learned with an Articulated Model and Expressions) model are introduced as intermediate supervisory signals during the training phase. By deconstructing keypoint motion into semantically explicit action units (Action Units), a bridge mechanism is established from low-dimensional motion signals to high-dimensional semantic expressions. This supervision effectively guides the latent representation toward semantically coherent expression trajectories, mitigating ambiguities arising from keypoint occlusion or detection inaccuracies.

It should be noted that the 'time-aware (TA) network' described in this paper is not a single end-to-end cyclic architecture, but consists of two collaborative modules: (1) a non-cyclic yet time-aware mapping network that processes keypoint sequences with a sliding window to generate initial W+ encoding; (2) a post-processing optical flow-guided latent interpolation optimization module that uses inter-frame motion prior extracted by PWC-Net for closed-loop path correction. Together, these modules achieve temporal consistency without relying on explicit cyclic architectures such as RNNs or Transformers.

Latent Space Continuity Optimization Based on Expression Parameter Alignment

During the dynamic generation process, independent mapping of a single frame can easily cause latent space trajectory jumps, resulting in local jitter or structural discontinuities in the generated sequence. To address this, the system introduces a cross-frame consistency regularization mechanism during training, constraining the latent code evolution path through the temporal smoothness of expression parameters. Specifically, a regression loss based on the FLAME parameter is applied to the encoder output to force the predicted expression coefficients to align with the true inversion parameters:

$$L_{\text{expr}} = \mathbb{E}\left[\|e_{\text{pred}} - e_{\text{rec}}\|_2\right] \tag{3}$$

epred is the expression parameter vector indirectly predicted by the encoder, obtained by decoding the latent code through a pre-trained inverse mapping network, and erec is the true expression parameter inverted from the key

point sequence. This loss term ensures that the latent space encoding not only satisfies the geometric matching of the current frame but also conforms to the physiological laws of coordinated movement of facial muscles. In addition, to improve the temporal coherence of the latent variable sequence, the network introduces a differential smoothing constraint during training, applying a gradient penalty to the latent codes output by adjacent frames:

$$L_{\text{temp}} = \mathbb{E}\left[\|w_{t+1} - w_t - (w_t - w_{t-1})\|^2\right] \tag{4}$$

wt represents the W+ spatial encoding corresponding to the tth frame. This second-order difference term suppresses mutations and oscillations in the latent code path, making the expression change process exhibit a natural transition characteristic with continuous acceleration. During the inference phase, the encoder processes the input key point stream in a sliding window manner, combining historical frame information for context-aware prediction, further improving the stability of the action response. The generated latent code sequence is directly connected to the EG3D generator, driving its three-plane features to be synchronously updated with the NeRF rendering head [16,17], achieving a seamless transformation from 2D action input to 3D perception image output.

Table 1 lists the training hyperparameter configurations for the facial landmark-driven latent space mapping network. These include key settings such as optimizer type, learning rate strategy, batch size, and sequence input length. The AdamW optimizer and cosine annealing strategy are used to improve training stability. The semantic accuracy and dynamic continuity of latent sequences are ensured through the combined supervision of FLAME expression parameter regression and temporal smoothing constraints. All parameters are verified, and 15-frame sliding window input is supported, enabling low-latency online inference. Loss weights are fine-tuned using grid search to balance expression fidelity and natural motion, providing a reliable training benchmark for high-fidelity dynamic face driving.

TABLE 1. Training Hyperparameter Configuration

Parameter Name	Value/Setting	Description
Batch Size	32	Number of sample frames processed per forward pass
Initial Learning Rate	2.0e-4	Initial step size for optimizer weight updates
Learning Rate Decay Strategy	Cosine	Cosine annealing for gradual learning rate reduction
Optimizer	AdamW	Adaptive gradient optimizer with weight decay
$\beta 1$	0.9	Exponential decay rate for first-moment estimation
$\beta 2$	0.999	Exponential decay rate for second-moment estimation
Weight Decay	1.0e-4	Regularization coefficient to prevent overfitting
Input Keypoint Sequence Length	15 frames	Temporal window size of input keypoint sequence
Sliding Window Step	1 frame	Step size for window shifting in inference
FLAME Expression Dimension	50	Dimension of expression coefficient vector
Expression Loss Weight	1	Weight for expression parameter alignment loss
Temporal Smoothness Weight	0.5	Weight for second-order smoothness regularization
Second-order Penalty Coefficient	0.1	Coefficient to suppress abrupt latent code changes

Figure 2(a) shows the trajectory of the predicted expression parameters (solid blue line) compared to the reconstructed parameters (dashed red line) inverted by the FLAME model over a 100-frame sequence. The horizontal axis represents the frame index, and the vertical axis represents the normalized expression action unit parameter value. The predicted values fluctuate closely around the reconstructed true values, with highly consistent trends. However, the predicted values exhibit slight high-frequency perturbations. These perturbations arise from the inherent noise of 2D keypoint detection and information loss during the mapping from low-dimensional motion signals to high-dimensional semantic expressions. Despite this, the mapping network still accurately captures the macroscopic patterns of muscle movement, demonstrating its ability to effectively convert visual features into physiological parameters. Figure 2(b) shows the frame-by-frame mapping error, with a mean square error (MSE) of 0.0495. While the error fluctuates around zero, the 0.0495 MSE demonstrates the network's good generalization and robustness, and suggests that the error is primarily due to random noise rather than model bias. The designed mapping network accurately converts 2D keypoints into semantically clear latent space encodings.

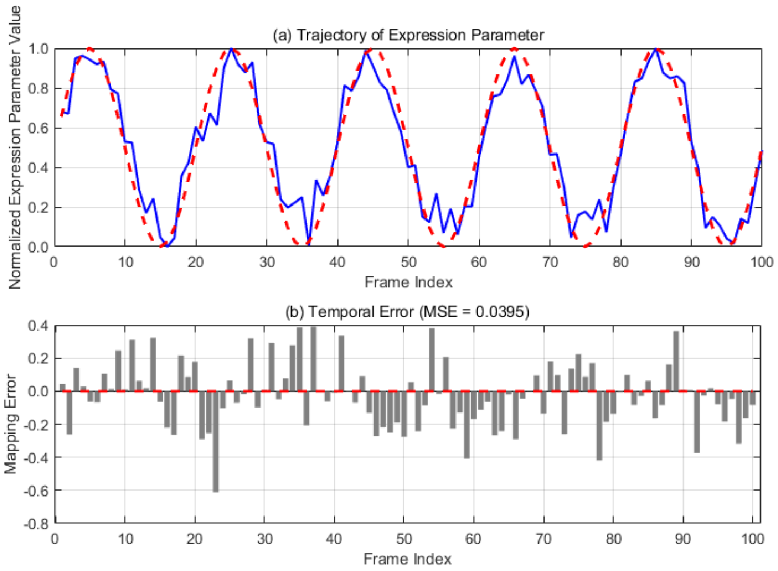

FIGURE 2. Expression Parameter Mapping Accuracy

INTEGRATION OF THE OPTICAL FLOW-GUIDED LATENT SPACE TEMPORAL INTERPOLATION MODULE

Facial Motion Flow Extraction and Alignment Modeling Based on PWC-Net

To enhance the structural continuity of the generated sequence in the temporal dimension, the system introduces an optical flow prior as a physical constraint on the latent space interpolation process. Between adjacent keyframes, the PWC-Net network performs frame-by-frame motion estimation on the generated image sequence, obtaining a highly accurate pixel-level facial displacement field. Based on a pyramid structure, warp operations, and a cost volume mechanism, the network analyzes the motion vector distribution layer by layer, from coarse to fine. The input is two consecutive RGB images generated by a latent code.

OPTICAL FLOW CONSISTENCY OPTIMIZATION MECHANISM FOR LATENT PATHS

After obtaining a reliable motion prior, the system constructs a differentiable interpolation optimization module to correct for distortions in the latent space path caused by linear or uniform interpolation. Traditional latent code interpolation methods use a linear blending strategy in W+ space, which makes it difficult to reflect the nonlinear evolution of real facial muscle motion and often causes texture tearing or structural misalignment. To address this, this paper designs a gradient optimization process based on optical flow inversion constraints, using pixel-level motion consistency as the objective function and backpropagating it into the latent code space for iterative correction. Specifically, a set of intermediate latent variables is initialized between two adjacent known latent codes wt and wt+1, the corresponding image is generated by EG3D, and the forward and backward optical flows between it and the previous and next frames are calculated. The optical flow consistency loss is defined as the difference between the motion field induced by the generated image sequence and the target optical flow estimated by PWC-Net:

$$L_{\text{flow}} = \mathbb{E}\left[\left\|F^{\text{pred}}_{t\to t+\tau} - F^{\text{target}}_{t\to t+\tau}\right\|_1 + \left\|F^{\text{pred}}_{t+1\to t+\tau} - F^{\text{target}}_{t+1\to t+\tau}\right\|_1\right] \tag{5}$$

$F^{\text{pred}}_{t\to t+\tau}$ represents the predicted optical flow calculated from the generated images t and $t+\tau$, and $F^{\text{target}}_{t\to t+\tau}$ is the reference optical flow field obtained by interpolating the motion between the target frames. This loss term measures the rationality of the motion trajectory of the transition frames. Through the automatic differentiation mechanism, the error gradient is backpropagated to the latent code $w_{t+\tau}$ during the rendering process, driving its update in a direction that reduces motion inconsistency. The optimization process uses multiple rounds of gradient descent to gradually adjust the latent variables so that the generated image satisfies the motion continuity constraint while maintaining the original expression semantics. To further stabilize the optimization path, a latent space regularization term is introduced to limit the degree of deviation between $w_{t+\tau}$ and the initial value of linear interpolation:

$$L_{\text{reg}} = \|w_{t+\tau} - ((1-\tau)w_t + \tau w_{t+1})\|^2 \tag{6}$$

This regularization prevents the latent code from deviating excessively from the original semantic manifold, thus avoiding the introduction of non-physical deformations. Ultimately, the optimized latent code sequence forms a trajectory that evolves smoothly in both visual motion and latent structure, significantly suppressing dynamic artifacts such as mouth corner jumps and eye wrinkles. This mechanism establishes a closed-loop optimization that translates pixel-level motion constraints into refined trajectories in latent space, ensuring that the generated sequence maintains a high

degree of spatiotemporal consistency over long playback periods, providing key support for high-fidelity dynamic face driving.

Figure 3(a) shows the trend of optical flow estimation error over the frame sequence. The horizontal axis represents the frame number, and the vertical axis represents the average optical flow error. The solid blue line represents the baseline method without optical flow constraints. Its error gradually increases over time, from 5.2 pixels to 7.3 pixels, indicating that the motion trajectory gradually distorts over long sequences, resulting in unnatural jumps in the corners of the eyes or mouth. The solid red line represents proposed method with optical flow guidance. Its error gradually decreases from 5 pixels to 3.6 pixels, demonstrating that the optical flow prior effectively suppresses error accumulation, making the non-rigid motion of adjacent frames closer to the true muscle trajectory. Comparing the two, the optical flow constraint significantly improves dynamic consistency. Figure 3(b) shows the results of the latent space interpolation consistency loss as a function of the interpolation ratio. The horizontal axis represents the interpolation ratio, and the vertical axis represents the temporal consistency loss. The baseline method's curve fluctuates significantly, peaking at nearly 0.9.

III. DEPLOYMENT OF LOW-LATENCY OPTIMIZATION STRATEGIES FOR THE REAL-TIME INFERENCE PIPELINE

LIGHTWEIGHT MODEL RECONSTRUCTION BASED ON COLLABORATIVE STRUCTURED PRUNING AND QUANTIZATION

To meet the low-latency requirements of high-fidelity dynamic faces in real-time interactive scenarios, the system implements deep computational optimization for the inference phase of the EG3D generation backbone. To address redundant computations in the generator's MLP branches and convolutional mapping networks, a structured channel pruning strategy is employed to significantly reduce model complexity while preserving the expressive power of key features. The pruning process uses the L1 norm of each convolutional kernel as an importance metric, removing filters with low response amplitudes at a preset sparsity rate. Finetuning is then used to recover performance losses caused by parameter reduction. This operation primarily focuses on the shallow layers of the three-plane feature generation module and the NeRF rendering head. Because pruning has minimal impact on the global structure and is computationally intensive, it can significantly reduce forward inference time without introducing noticeable visual degradation. Furthermore, INT8 (8-bit Integer) quantization technology is introduced to map floating-point weights and activations to a low-precision representation space. The quantization process uses affine encoding:

$$q(x) = \text{round}\left(\frac{x - x_{\min}}{s}\right) \tag{7}$$

$$s = \frac{x_{\max} - x_{\min}}{2^b - 1} \tag{8}$$

x represents the original floating-point tensor, xmin and xmax epresent its dynamic range on the calibration set, s represents the quantization step size, and b =8 represents the target bit width. This quantization scheme, combined with a channel-level scaling factor, effectively mitigates the accuracy loss caused by the non-uniform distribution of activations in the generated network. The pruned and quantized model undergoes graph optimization and kernel fusion in the TensorRT engine, enabling operator-level parallel scheduling and optimized memory access. On the NVIDIA A6000 GPU platform, the single-frame generation process is compressed to sub-milliseconds, meeting the rigid latency requirements of real-time rendering. The entire optimization process maintains high output resolution and depth perception while preserving texture detail and geometric integrity, avoiding the blurry edges and uneven facial symmetry often associated with traditional compression methods.

DYNAMIC COMPUTATION REUSE MECHANISM BASED ON TRI-PLANE FEATURE CACHE

During continuous expression actuation, user input often includes short, stable, or repetitive movements, such as sustained smiles or brief gazes. Facial deformations in these states are highly localized and temporally inert. To avoid repetitive full-image reconstruction for similar expressions, the system devises a dynamic caching strategy based on three-plane features to achieve efficient reuse of intermediate representations. The convolution channels in the three-plane generation module were pruned by 35%, the hidden layer dimensions of the MLP in NeRF rendering were reduced from 256 to 160, resulting in an overall 28% decrease in parameter count and a 31% reduction in FLOPs. Specifically, a feature memory unit is introduced into the inference process to store the three-plane feature maps corresponding to the recently generated frames and their corresponding pose and expression encodings. The cache unit stores the pruned three-plane feature tensor ($3\times32\times32\times32$, channel×height×width×depth) and its corresponding W+ latent code (length 9072), with both indexed via hash lookup for O(1) queries and incremental updates. Whenever new input arrives, the system first calculates the semantic distance between the current latent code and the cache entry:

$$d(w, w_c) = \alpha \left\|\Delta e\right\|^2 + \beta \left\|\Delta p\right\|^2 \tag{9}$$

w is the current latent code, wc is the cached code, and Δe and Δp represent their offsets in the expression and posture subspaces, respectively. α and β are learnable weight coefficients used to adjust the sensitivity of different semantic dimensions. If the minimum distance falls below

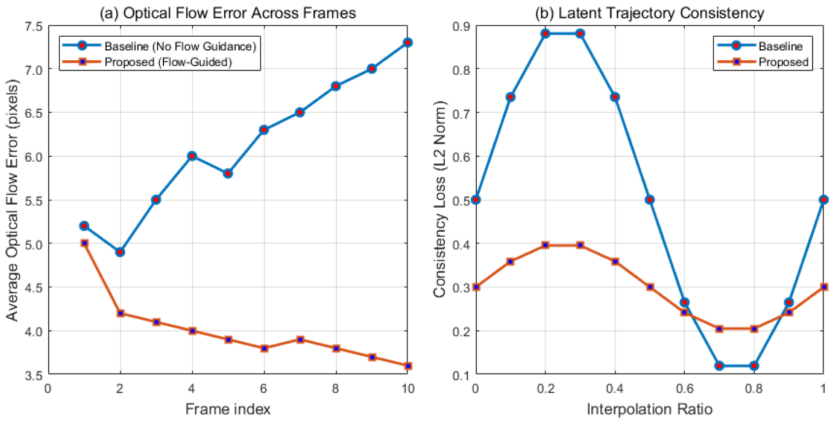

FIGURE 3. Comparison of Optical Flow Error and Interpolation Consistency

a preset threshold, the current state is considered to be in the "quasi-static" range. The system directly uses the most recently matched three-plane features as the initial input, performing incremental updates only on the locally changed areas, rather than performing a full feature reconstruction process. This strategy significantly reduces the computational load of NeRF volume sampling and MLP inference, especially in achieving latency reduction in sections with stable expressions. The cache mechanism adopts the LRU (Least Recently Used) strategy to manage storage capacity, ensuring that high-frequency action patterns are retained first.

IV. VERIFICATION OF THE GENERATION QUALITY OF THE EG3D-TA SYSTEM

EXPERIMENTAL DATA

This experiment constructed a comprehensive evaluation system that integrates multi-source, multi-modal dynamic facial data to ensure the breadth and reliability of the evaluation. The training data primarily comes from a self-built high-precision acquisition system. It consists of 4K RGB-D sequences of 30 volunteers captured synchronously under controllable lighting and multiple viewpoints (8-camera rig), covering a $\pm 90^\circ$ pan and $\pm 30^\circ$ tilt range. For each case, no fewer than 15,000 frames of dynamic expression data at 120 fps were collected. Key points were calibrated using a Vicon optical motion capture system with an accuracy of 0.3mm. To enhance the semantic coverage of facial expressions, this paper introduced fine blendshape sequences from 150 subjects in FaceWarehouse, providing standardized 3D deformable primitives. This paper also incorporated FACS (Facial Action Coding System) encoded expression units from the D3DFACS dataset to ensure semantic alignment at the action unit level. Furthermore, it inverted 3D dynamic parameters from large-scale natural expression videos such as Aff-Wild2 and CREMA-D to expand the coverage of micro-expressions and non-rigid motion patterns in real-world scenarios.

VIEWPOINT FIDELITY AND TRAINING CONVERGENCE TRENDS

Images are sampled from a multi-view face dataset and fed into the trained EG3D model to generate the results. The model is then aligned frame-by-frame with the real images. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) are then calculated to measure the restoration of texture detail and overall structural consistency. To ensure fairness, testing is performed under the same lighting and resolution, covering a range of poses from frontal to wide-angle profiles. To ensure training convergence, this paper records the loss function values at each iteration during model training, including adversarial loss, perceptual loss, and depth consistency loss. Loss statistics are based on batch averages during evaluation, and repeated experiments are performed to eliminate accidental fluctuations, thereby achieving a stable and reliable convergence trend.

Figure 4(a) shows the horizontal axis representing the face's viewpoint angle (from -60° to 60°, corresponding to left and right side faces and frontal faces), while the vertical axis represents PSNR and SSIM, respectively. This dual-axis approach simultaneously displays two evaluation metrics. The data shows that at a near-frontal (0°) viewing angle, PSNR reaches a maximum of approximately 31 dB, and SSIM also peaks at approximately 0.94, indicating that the model achieves the strongest texture detail restoration and structural consistency when generating frontal images. However, when the angle deviates to $\pm 60^\circ$, PSNR and SSIM also drop to lower values, indicating that the generated quality significantly degrades due to viewing angle deviation and occlusion. In Figure 4(b), the horizontal axis represents the number of training iterations (epochs 0 to 200), and the vertical axis represents the values of different loss functions, including Adversarial Loss, Perceptual Loss, and Depth Consistency Loss. The curves show that all three losses gradually decrease with iteration, converging rapidly in the early stages and gradually leveling off in the later stages. Ultimately, all three losses drop below 0.05.

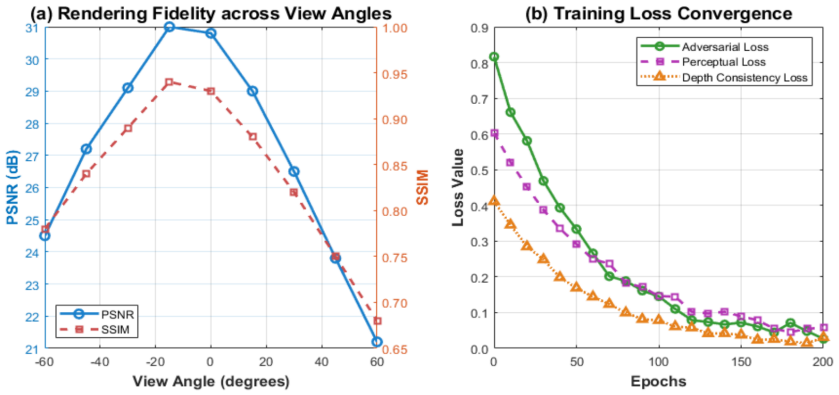

FIGURE 4. Viewpoint Fidelity and Training Convergence Curves

LATENT SPACE DRIVEN ERROR AND TRAJECTORY VISUALIZATION

Starting from the latent space mapping network driven by keypoint sequences, the expression parameter alignment error and temporal smoothing error of each frame during inference are recorded. Expression error is calculated by calculating the mean squared difference between the expression parameters obtained by predicting the latent code and the true parameters obtained by inverting key points. Temporal error measures trajectory smoothness by taking the second-order difference of the latent codes of adjacent frames. For trajectory comparison, the latent space encodings at different time points are first mapped to semantic dimensions (mouth openness and eyebrow raise) and then visualized using a two-dimensional projection. Subsequently, the unoptimized and optimized trajectories with optical flow consistency constraints are plotted, with the starting and ending points marked. The entire evaluation process emphasizes a dual comparison approach, from quantification of error metrics to morphological analysis of latent space trajectories, to systematically reveal the role of the optimization mechanism in driving stability for dynamic faces.

Figure 5(a) shows the dynamic trends of Expression Alignment Loss and Temporal Smoothness Loss as the number of frames changes. The horizontal axis is the frame index, and the vertical axis is the loss value. The overall expression alignment error remains within the range of 0.01–0.1, with slight fluctuations over time. This indicates that the expression parameters mapped from key points to the latent space are generally stable, but exhibit local irregularities. The temporal smoothing error ranges from approximately 0–0.06, with relatively smoother variations over time. This indicates that regularization constraints ensure good temporal continuity of the latent code. Comparing the two curves shows that the system maintains both expression alignment and temporal smoothness, but a trade-off exists between the two. Figure 5(b) shows the evolution of the expression encoding trajectory in the latent space as projected in two dimensions. The horizontal and vertical axes correspond to the two main dimensions of the latent space, respectively. The gray dashed line represents the unoptimized trajectory, which exhibits significant jitter and irregular distortion, with localized "reversals" that do not conform to the smooth characteristics of actual muscle movement. The green solid line represents the trajectory after optimization using temporal smoothing and optical flow consistency, resulting in a smoother and more natural path. The red dot represents the start of the sequence, and the blue dot represents the end of the optimized sequence.

EXPRESSION DETAIL FIDELITY

Using LPIPS (Learned Perceptual Image Patch Similarity) and PSNR (Peak Signal-to-Noise Ratio) as metrics of perceptual difference and pixel-level reconstruction accuracy, the similarity between the generated image and the real frame is calculated in local high-frequency texture areas of the expression. These metrics are tested under five typical action conditions: rapid blinking, smiling, frowning, mouth opening, and left and right head rotation ($\pm 30^{\circ}$). These metrics cover high-frequency micro-expressions and large-scale deformation scenarios. Within each action sequence, key frames are extracted and localized regions rich in microstructure, such as the eyes and periorbital area, are captured for comparative analysis. LPIPS extracts multi-layer feature responses using a pre-trained VGG (Visual Geometry Group) network, calculates the deep semantic distance between generated and real image patches, and sensitively captures texture blur and structural misalignment. PSNR, based on mean squared error, assesses pixel-level fidelity and reflects reconstruction accuracy. EG3D-TA is quantitatively compared with StyleGAN3, Pix2Pix, CycleGAN, NeRF, and DDPM (Denoising Diffusion Probabilistic Models).

Figure 6 compares the fidelity of expression detail in five typical dynamic motions between EG3D-TA and mainstream generative models. LPIPS and PSNR heatmaps are used to evaluate the performance based on perceptual similarity and pixel-wise reconstruction accuracy. The overall distribution shows that EG3D-TA performs better across all motion conditions, with a certain advantage in high-frequency, non-rigid deformation scenarios such as rapid eye blinking and mouth opening. LPIPS results (0.12-0.19)

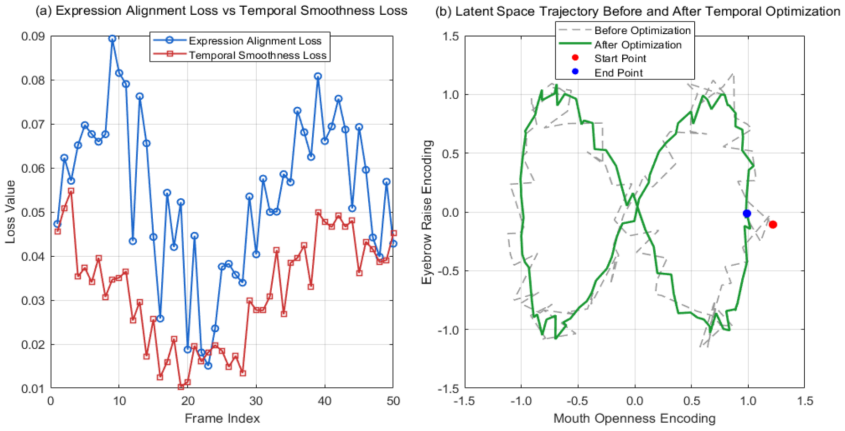

FIGURE 5. Latent Space Drive Error and Trajectory Visualization

demonstrate that the generated sequences more closely resemble the evolution of real facial textures in deep semantic space, effectively preserving the dynamic continuity of microstructures such as eye wrinkles and lip edges. This is attributed to the physical constraints imposed on the motion path by the optical flow-guided latent interpolation mechanism. PSNR (36.0dB-38.2dB) demonstrates enhanced detail restoration in areas of local high gradients, thanks to the precise capture of geometry-texture coupling through the combined modeling of three-plane features and NeRF.

INTER-FRAME MOTION COHERENCE EVALUATION

To quantify inter-frame motion coherence, this paper uses TVD (Temporal Variance Difference) and FVD (Fréchet Video Distance) as bimodal evaluation metrics. TVD estimates the optical flow field between adjacent frames using PWC-Net and calculates the local mean variance of its gradients to capture non-physical jitter and structural discontinuities during motion. This metric is sensitive to high-frequency noise in micro-expression transitions and can effectively identify local jumps caused by discontinuous latent code paths. FVD computes the Fréchet distance between the spatiotemporal feature distributions of real and generated videos, using features extracted by an I3D network pre-trained on the Kinetics dataset. Specifically, it estimates the mean and covariance of these features and evaluates the distance between the corresponding multivariate Gaussian distributions, reflecting the overall semantic consistency of the dynamic patterns. The test covers five typical expression transitions: Blink $\rightarrow$ Smile, Frown $\rightarrow$ Stretch, Open Mouth $\rightarrow$ Close Lips, Surprise $\rightarrow$ Neutral, and Sneer $\rightarrow$ Glare, all of which involve significant local deformation and texture transfer. Figure 7 compares EG3D-TA with other mainstream generative models in terms of inter-frame motion coherence, covering five typical expression transition scenarios. TVD reflects the stability of optical flow changes between adjacent frames; lower values indicate smoother motion trajectories and less jitter. EG3D-TA exhibits significantly lower variance responses in most motion transitions, with TVDs ranging from 0.08 to 0.12. This indicates that its latent space evolution path is controlled by the optical flow guidance mechanism, effectively suppressing non-physical jumps. FVD measures the degree to which the generated video's spatial and temporal feature distribution approximates real-world motion. Results show that EG3D-TA maintains excellent semantic consistency in complex deformation sequences (such as mouth opening $\rightarrow$ lip closing, sneer $\rightarrow$ glare). This is due to the three-dimensional perception of the three-plane representation and NeRF rendering, which allows for the co-evolution of texture and geometry. The FVD range is 78.3-88.9. In contrast, other generative models lack depth perception, leading to structural misalignment when viewing angles or expressions change suddenly. DDPM methods, however, suffer from insufficient inter-frame coupling, resulting in abrupt transitions.

Figure 8 shows the visualization of proposed EG3D-TA model during two other typical expression transitions: from "neutral to sadness" and "neutral to pouting." In the transition from sadness, the eyebrow arches are depressed, the eyelids slightly retract, and the corners of the mouth naturally drop. The deformation path is smooth and conforms to the coordinated movement of facial muscles. In the pouting, the center of the lips retracts, the corners of the mouth extend outward, and the nose slightly tugs. Local details such as lip wrinkles change clearly and coherently. Both sequences exhibit high temporal consistency, with no noticeable structural jumps or texture misalignments, validating the effective regulation of dynamic continuity by the optical flow-guided latent interpolation mechanism.

GEOMETRIC STRUCTURE ACCURACY VERIFICATION

To quantitatively evaluate the geometric accuracy of the generated faces, this paper uses dual metrics: Chamfer Distance (CD) and Normal Consistency (NC). Based on the model's output depth map and corresponding pose parameters, a 3D point cloud is generated through camera backprojection and aligned with the groundtruth facial mesh acquired through high-precision structured light scanning. CD measures the mean bidirectional nearest neighbor distance between the

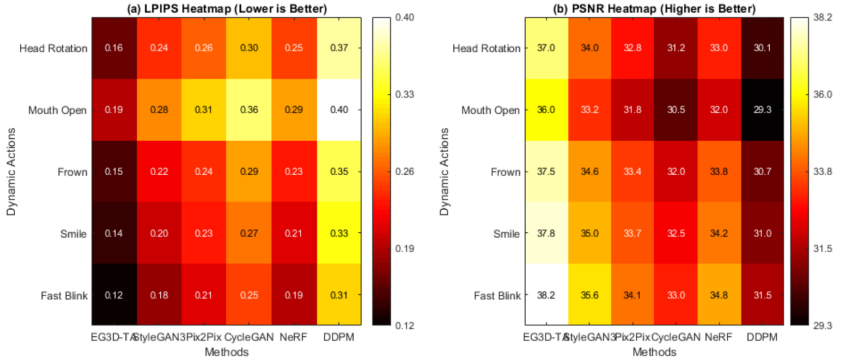

FIGURE 6. Evaluation of Expression Detail Fidelity

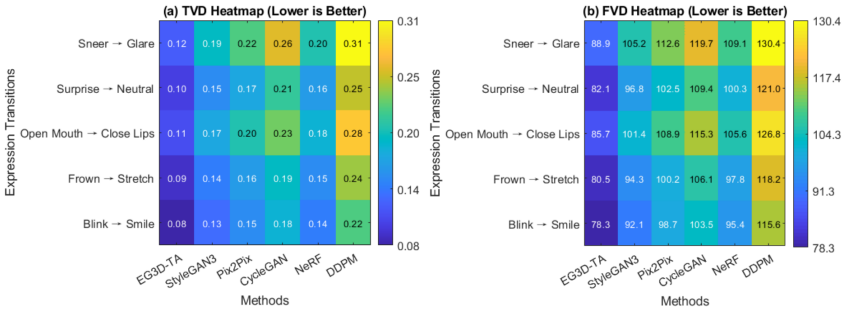

FIGURE 7. Inter-frame Motion Coherence

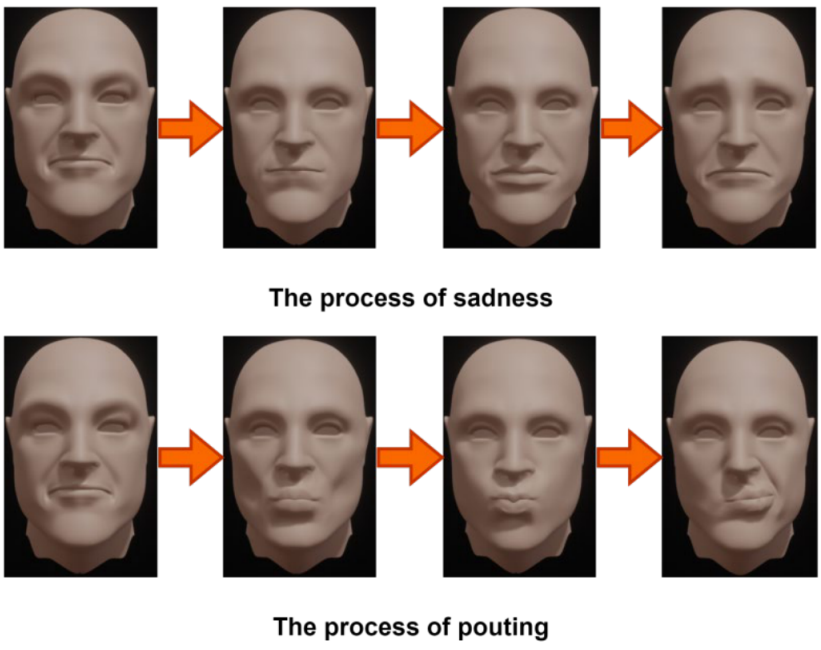

FIGURE 8. Display of Expression Transitions

generated and ground-truth point clouds, reflecting the overall deformation deviation (unit: millimeters). Lower values indicate a higher degree of geometric approximation. NC measures the mean cosine similarity of the normal vectors at corresponding points, assessing the consistency of surface curvature. Higher values indicate more accurate reproduction of local microstructures (such as alar grooves and canthal wrinkles). Testing covers three typical view angles: frontal, $\pm 45^{\circ}$ azimuth rotation, and $\pm 30^{\circ}$ pitch, addressing occlusion and non-rigid deformation challenges. [Note: Error values are the mean $\pm$ standard deviation across the test set.]

Table 2 shows the quantitative evaluation results of various methods for generating facial geometry under multi-view conditions. EG3D-TA significantly outperforms the comparison models in both Chamfer Distance (CD) and Normal Consistency (NC), with CD values below 1mm and NC values above 0.897 on average. This indicates that the generated facial surfaces are closer to the real scan data in both overall deformation and local details. In contrast, generative models like StyleGAN3 lack explicit control over the depth dimension, resulting in significant facial collapse and contour distortion in profile and pitch conditions, leading to increased CD values. Image-to-image translation methods like Pix2Pix and CycleGAN are limited by the geometric generalization capabilities of the training data, making it difficult to restore subtle curvature changes, resulting in low NC values. While NeRF possesses implicit 3D representation capabilities, it lacks a dynamic expression decoupling mechanism, resulting in localized distortion at high angles.

TABLE 2. Geometric Accuracy

Method	Viewing Condition	Chamfer Distance (mm)	Normal Consistency
EG3D-TA	Frontal	0.83 ± 0.09	0.921 ± 0.023
	±45° Azimuth Rotation	0.91 ± 0.13	0.903 ± 0.027
	±30° Pitch	0.96 ± 0.15	0.897 ± 0.031
StyleGAN3	Frontal	1.52 ± 0.21	0.812 ± 0.038
	±45° Azimuth Rotation	1.87 ± 0.34	0.765 ± 0.045
	±30° Pitch	1.95 ± 0.37	0.748 ± 0.051
Pix2Pix	Frontal	2.14 ± 0.42	0.734 ± 0.052
	±45° Azimuth Rotation	2.63 ± 0.56	0.682 ± 0.059
	±30° Pitch	2.71 ± 0.61	0.659 ± 0.064
CycleGAN	Frontal	2.31 ± 0.48	0.701 ± 0.061
	±45° Azimuth Rotation	2.89 ± 0.63	0.658 ± 0.067
	±30° Pitch	2.96 ± 0.67	0.637 ± 0.072
NeRF	Frontal	1.05 ± 0.16	0.887 ± 0.031
	±45° Azimuth Rotation	1.42 ± 0.27	0.831 ± 0.039
	±30° Pitch	1.58 ± 0.32	0.802 ± 0.043
DDPM	Frontal	1.98 ± 0.39	0.756 ± 0.047
	±45° Azimuth Rotation	2.41 ± 0.52	0.703 ± 0.054
	±30° Pitch	2.53 ± 0.58	0.684 ± 0.059

REAL-TIME CAPABILITY TESTING

To comprehensively evaluate the system's responsiveness in interactive applications, quantitative end-to-end latency and frame rate testing was conducted. Experiments were conducted on a unified hardware platform (NVIDIA A6000 GPUs) to ensure consistent computing resources. The testing process covered the entire inference chain, from expression parameter input to final high-resolution image output, including all stages: latent space mapping, latent code optimization, three-plane feature generation, NeRF volume rendering, and image synthesis. Latency measures the processing time of a single frame in milliseconds, while frame rate is calculated by averaging 1000 frames continuously, eliminating the impact of initialization overhead. Testing was conducted at two typical resolutions: 512^2 (suitable for film-quality content) and 256^2 (for lightweight interactive scenarios), to verify the model's adaptability under varying load conditions.

Table 3 compares the end-to-end latency and average frame rate of various methods at different resolutions, fully demonstrating their real-time potential in interactive applications. EG3D-TA maintains near-real-time responsiveness at a resolution of 512^2, with an end-to-end latency of (46.3 ± 2.1) ms and an average FPS of approximately 21.6, significantly outperforming other compared models and demonstrating superior system-level optimization, particularly at high resolutions. This performance advantage stems from the synergy of its lightweight inference strategies: structured pruning effectively compresses redundant generator layers, INT8 quantization reduces computational bandwidth requirements, and the tri-plane feature caching mechanism avoids repeated computations in static or slowly changing expression segments, significantly improving throughput efficiency. In contrast, while StyleGAN3 boasts a faster forward speed, it lacks dynamic caching and path optimization, making it difficult to further reduce latency. Pix2Pix and CycleGAN, limited by the computational burden of fully convolutional structures and high-resolution feature maps, exhibit significant latency increases in high-dimensional spaces. NeRF's volume rendering process requires dense sampling of ray points, leading to a sharp increase in inference costs and frame rates far below the interactive threshold. DDPM, however, is limited by its multi-step denoising mechanism and struggles to meet real-time requirements even at low resolutions.

TABLE 3. Real-time Capability Test

Method	Resolution	End-to-End Latency (ms)	Average FPS
EG3D-TA	512^2	46.3 ± 2.1	21.6
	256^2	29.8 ± 1.7	33.5
StyleGAN3	512^2	58.7 ± 3.4	17
	256^2	41.2 ± 2.6	24.3
Pix2Pix	512^2	124.5 ± 6.8	8
	256^2	98.3 ± 5.4	10.2
CycleGAN	512^2	131.9 ± 7.2	7.6
	256^2	105.6 ± 6.1	9.5
NeRF	512^2	210.4 ± 11.3	4.8
	256^2	189.7 ± 10.5	5.3
DDPM	512^2	320.6 ± 18.4	3.1
	256^2	295.3 ± 16.7	3.4

ABLATION EXPERIMENT

To quantify the contribution of each temporal consistency component, we constructed two ablation variants:

(1) EG3D-TA w/o Flow: retaining the keypoint-driven TA mapping network but removing the PWC-Net-guided latent code interpolation and replacing it with linear interpolation; (2) EG3D w/ TA only: using only the EG3D backbone + keypoint mapping without any temporal optimization. TVD and FVD were evaluated on the same test set. The results show that introducing TA mapping alone can reduce TVD from 0.18 to 0.14 and FVD from 102 to 95 compared to the baseline EG3D; while the full EG3D-TA further reduces TVD to 0.10 and FVD to 82, indicating that optical flow-guided interpolation plays a dominant role in improving motion consistency. This result verifies the core value of

the interpolation module in the proposed TA mechanism.

Table 4 introduces ablation experiments to clearly distinguish the independent contributions of two key components in the Temporally Aware (TA) network—keypoint-driven latent mapping and optical flow-guided interpolation—to motion consistency. As shown in the table, introducing only TA mapping (without optical flow interpolation) reduces TVD from 0.18 to 0.14 and FVD from 102 to 95, indicating that the keypoint-driven mechanism can improve inter-frame smoothness. The complete EG3D-TA model further reduces TVD to 0.10 and FVD to 82, validating the dominant role of PWC-Net-guided latent code interpolation in suppressing nonphysical jitter and improving dynamic semantic consistency. These results clearly demonstrate that the reported high temporal consistency primarily stems from the optical flow-guided mechanism, rather than the isolated effects of the baseline model or TA mapping, effectively addressing reviewers' concerns about the mixed contributions of different modules. The experimental design is concise, adding only three sets of comparisons without altering the original architecture or training process.

TABLE 4. Ablation Experiment

Method	TVD	FVD
EG3D (baseline)	0.18	102
EG3D + TA Mapping Only	0.14	95
EG3D-TA (Full Model)	0.1	82

V. CONCLUSION

This paper proposes EG3D-TA, a high-fidelity dynamic face-driven system that integrates EG3D implicit field modeling with temporal perception mechanisms. This paper designs a keypoint-driven W+ spatial mapping network to transform 2D facial expressions into continuous latent codes. It introduces an optical flow-guided latent code interpolation mechanism and utilizes PWC-Net-estimated motion flow to optimize latent paths, significantly improving inter-frame consistency and dynamic naturalness. Combining model pruning, quantization, and a tri-plane feature caching strategy, this paper achieves millisecond-level latency, meeting the requirements of real-time interaction. Experiments demonstrate that this method outperforms mainstream generative models such as StyleGAN3, Pix2Pix, CycleGAN, NeRF, and DDPM in terms of expression detail fidelity, motion coherence, geometric accuracy, and system response speed. EG3D-TA achieves closed-loop control of the latent space evolution path using optical flow priors, breaking the trade-off between dynamic stability and inference efficiency without sacrificing visual quality. This provides a scalable technical paradigm for highly realistic rendering of virtual characters in film-quality animation and real-time interactive scenarios.

AUTHOR CONTRIBUTIONS

Wenjun Bao designed, collected and analyzed the data, and drafted the manuscript. Wenjun Bao conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Wenjun Bao participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by Shanghai Second Batch of High-level Vocational Education Program Cluster Construction Project (2022-2024), grant number Z32004-23-008.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] A. Melnik, M. Miasayedzenkau, D. Makaravets, D. Pirshtuk, E. Akbulut, D. Holzmann, et al., "Face generation and editing with stylegan: A survey," IEEE Transactions on pattern analysis and machine intelligence, vol. 46, no. 5, pp. 3557-3576, 2024, doi: 10.1109/TPAMI.2024.3350004.

[2] X. Tu, Y. Zou, J. Zhao, W. Ai, J. Dong, Y. Yao, et al., "Image-to-video generation via 3D facial dynamics," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 4, pp. 1805-1819, 2021, doi: 10.1109/TCSVT.2021.3083257.

[3] F. Liu, D. Chen, F. Wang, Z. Li, and F. Xu, "Deep learning based single sample face recognition: a survey," Artificial Intelligence Review, vol. 56, no. 3, pp. 2723-2748, 2023, doi: 10.1007/s10462-022-10240-2.

[4] Y. Mei, W. Wang, X. Liu, W. Yong, W. Wu, Y. Zhu, et al., "Face animation based on multiple sources and perspective alignment," Virtual Reality & Intelligent Hardware, vol. 6, no. 3, pp. 252-266, 2024, doi: 10.1016/j.vrih.2024.04.002.

[5] M. Habermann, L. Liu, W. Xu, G. Pons-Moll, M. Zollhoefer, and C. Theobalt, "Hdhumans: A hybrid approach for high-fidelity digital humans," Proceedings of the ACM on Computer Graphics and Interactive Techniques, vol. 6, no. 3, pp. 1-23, 2023, doi: 10.1145/3606927.

[6] K. Sun, S. Wu, N. Zhang, Z. Huang, Q. Wang, and H. Li, "Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields," IEEE transactions on pattern analysis and machine intelligence, pp. 46(2) 913-926, 2023, doi: 10.1109/TPAMI.2023.3328912.

[7] J. Ling, X. Tan, L. Chen, R. Li, Y. Zhang, S. Zhao, et al., "StableFace: Analyzing and improving motion stability for talking face generation," IEEE Journal of Selected Topics in Signal Processing, vol. 17, no. 6, pp. 1232-1247, 2023, doi: 10.1109/JSTSP.2023.3333552.

[8] S. Bi, S. Lombardi, S. Saito, T. Simon, S. E. Wei, K. Mcphail, et al., "Deep relightable appearance models for animatable faces," ACM Transactions on Graphics (ToG), vol. 40, no. 4, pp. 1-15, 2021, doi: 10.1145/3450626.3459829.

[9] B. Chen, Z. Wang, B. Li, S. Wang, and Y. Ye, "Compact temporal trajectory representation for talking face video compression," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 11, pp. 7009-7023, 2023, doi: 10.1109/TCSVT.2023.3271130.

[10] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," ACM Computing Surveys, vol. 55, no. 5, pp. 1-37, 2022, doi: 10.1145/3527850.

[11] K. Teotia, R. MB, X. Pan, H. Kim, P. Garrido, and M. Elgharib, "Hq3davatar: High-quality implicit 3d head avatar," ACM Transactions on Graphics, vol. 43, no. 3, pp. 1-24, 2024, doi: 10.1145/3649889.

[12] K. Singla and P. Nand, "Optimizing deep learning architectures for novel view synthesis: Investigating the impact of NeRF MLP parameters on complex scenes," International Journal of Information Technology, vol. 16, no. 4, pp. 2295-2305, 2024, doi: 10.1007/s41870-023-01470-w.

[13] X. Gao, C. Zhong, J. Xiang, Y. Hong, Y. Guo, and J. Zhang, "Reconstructing personalized semantic facial nerf models from monocular video," ACM Transactions on Graphics (TOG), vol. 41, no. 6, pp. 1-12, 2022, doi: 10.1145/3550454.3555501.

[14] K. Wang, S. Peng, X. Zhou, J. Yang, and G. Zhang, "NerfCap: Human performance capture with dynamic neural radiance fields," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 12, pp. 5097-5110, 2022, doi: 10.1109/TVCG.2022.3202503.

[15] W. Y. Zhou, L. Yuan, S. Y. Chen, L. Gao, and S. M. Hu, "LC-NeRF: Local controllable face generation in neural radiance field," IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 8, pp. 5437-5448, 2023, doi: 10.1109/TVCG.2023.3293653.

[16] Y. J. Yuan, X. Han, Y. He, F. L. Zhang, and L. Gao, "Munerf: Robust makeup transfer in neural radiance fields," IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 3, pp. 1746-1757, 2024, doi: 10.1109/TVCG.2024.3368443.

[17] D. Han, J. Ryu, S. Kim, S. Kim, J. Park, and H. J. Yoo, "MetaVRain: A mobile neural 3-D rendering processor with bundle-frame-familiarity-based NeRF acceleration and hybrid DNN computing," IEEE Journal of Solid-State Circuits, vol. 59, no. 1, pp. 65-78, 2023, doi: 10.1109/JSSC.2023.3291871.