# Research on an AI Visual Effects Generation Framework for Virtual Digital Humans in Film and Television Production

**Peng Yan[1,*], Yubing Ma[2], Hailong Li[1], Qing Li[1], Jin Zhang[1]**
[1]School of Design and Art, Shandong Huayu University of Technology, Dezhou 253034, China
[2]School of Information Engineering, Shandong Huayu University of Technology,Dezhou 253034, China
Corresponding author: Peng Yan (e-mail: yp18613609980@163.com).

**ABSTRACT** The large-scale application of virtual digital humans in film and television has imposed extremely high demands on the generation precision and systematicity of visual effects. To address three major technical bottlenecks—high-fidelity 3D reconstruction of film-grade virtual digital humans, temporal driving consistency, and rendering realism in complex scenes—this paper presents a modular AI visual effects generation framework tailored to film and television production pipelines. The framework achieves collaborative geometry-texture modeling through joint optimization of NeRF and GAN, and resolves stage-wise representation inconsistency via cross-module feature alignment. Leveraging Transformer-based multi-scale temporal modeling and joint control of skeletal keypoints, high-precision motion and expression restoration is realized. A dual-path rendering pipeline combining physical rendering and diffusion models ensures cross-scene visual consistency. It also supports multi-level quality adaptation to meet the differentiated demands of preview, editing and final delivery in actual production workflows. Experiments demonstrate that the framework effectively improves core metrics of digital human production, balances computational efficiency and output quality, and demonstrates practical viability for industrial deployment in film and television.

**INDEX TERMS** Virtual digital human; Neural radiance field; Visual effects generation; Physical rendering; Temporal driving

## I. INTRODUCTION

WITH the deep integration of deep learning and computer graphics, virtual digital humans have become an integral component in the film and television production pipeline. Unlike gaming or real-time interactive scenarios, film and television production imposes more stringent technical specifications on virtual digital humans: frame-level motion synchronization accuracy, high-fidelity facial geometric reconstruction, controllable illumination decoupling, and cross-scene rendering consistency, which constitute the fundamental bottlenecks impeding the practical deployment of existing technologies. While current mainstream methods have achieved significant progress at the individual module level, systematic deficiencies remain in cross-module collaborative optimization, end-to-end pipeline integration, and industrial-grade rendering efficiency. Existing NeRF-based [1] reconstruction methods lack unified feature representation constraints with downstream driving modules, making it difficult to simultaneously guarantee reconstruction precision and driving consistency; the rendering layer also lacks adaptive enhancement mechanisms for complex illumination conditions. In light of these considerations, we systematically propose an AI visual effects generation framework encompassing a three-layer architecture of reconstruction, driving, and rendering, with cross-module feature alignment as the core design principle, to achieve reproducible and quantitatively optimizable production workflows while ensuring visual quality.

## II. VISUAL QUALITY CONSTRAINTS AND TECHNICAL POSITIONING OF FILM-GRADE VIRTUAL DIGITAL HUMAN VISUAL EFFECTS

### *A. VISUAL QUALITY CONSTRAINT SYSTEM FOR VIRTUAL DIGITAL HUMANS IN FILM AND TELEVISION PRODUCTION*

#### 1) Visual Quality Constraint System for Virtual Digital Humans in Film and Television Production

Film and television production manifests quality constraints on virtual digital humans as joint tightening across multiple physical dimensions, with systematic coupling rather than independence among dimensional indices. At the geometric level, facial mesh vertex error must be controlled within

0.3 mm to support credible restoration of microstructures such as wrinkles and pores under close-up shots; this metric derives from benchmark requirements established in theatrical-grade production by current mainstream multi-view scanning methods (e.g., FaceScape dataset evaluation protocols). At the temporal level, frame-level motion alignment error based on the 24 fps standard must not exceed $\pm 1$ frame, and the acceptable upper limit of lip movement deviation (LMD) is defined as within 3.0 mm in the evaluation protocols of methods such as GeneFace [2]. At the illumination level, material response consistency across scene lighting conditions requires the rendering pipeline to possess explicit illumination decoupling capability, with CIEDE2000 color difference standard $\Delta E$ below 2.5 to avoid subjectively perceptible color casts. At the output specification level, theatrical-grade delivery demands 4K or even 8K resolution, PSNR no less than 38 dB, and SSIM higher than 0.92. The core contradiction of the above constraints lies in the nonlinear growth of computational overhead accompanying improved geometric precision, and the latency bottleneck between high-resolution rendering and real-time illumination decoupling; therefore, system design must unify multidimensional constraints as coupled optimization objectives.

### B. TECHNICAL LIMITATION ANALYSIS OF EXISTING AI SPECIAL EFFECTS GENERATION METHODS

The limitations of current methods are rooted in feature drift caused by independent module optimization, rather than insufficient performance of individual modules. At the reconstruction level, NerFACE [3] parameterizes expressions as 76-dimensional vectors and jointly optimizes them with NeRF, achieving PSNR of 28.75 dB under training-set self-reenactment scenarios (as reported by IM Avatar [6]); however, its implicit geometric representation does not output structured features to downstream driving modules, causing skeletal binding to lose geometric anchors. GaussianAvatars [4] achieves controllable animation through rigged 3D Gaussians, yet obvious geometric artifacts remain in topologically complex regions such as the oral cavity interior. EnNeRFACE [7] and InstantAvatar [8] further extend NeRF-based facial reconstruction to dynamic scenarios, while GaussianHead [9] introduces learnable Gaussian derivation for high-fidelity head avatars; however, these methods similarly struggle with structured feature output for downstream driving modules in topologically complex regions. At the driving level, GeneFace++ [5] advances LMD to near 3.0 mm through pitch contour auxiliary features and temporal loss, yet its motion prediction module and NeRF renderer likewise lack unified intermediate feature constraints, and the cross-domain audio-driven lip jitter issue remains fundamentally unresolved. At the rendering level, existing PBR-based methods encode diffuse reflection, specular reflection, and environmental illumination into network weights, making color casts and shadow distortion difficult to avoid during cross-scene transfer. The common root of the above problems is that each module adopts independent loss function optimization, lacking explicit feature alignment constraints during stage transitions, causing errors to accumulate and amplify progressively along the pipeline.

### C. OVERALL FRAMEWORK ARCHITECTURE DESIGN PRINCIPLES AND CROSS-MODULE COLLABORATIVE MECHANISMS

The core design motivation of the framework stems from the targeted response to the above cross-module feature drift problem. Its key distinction from existing works such as NerFACE [3] and IM Avatar [6] is that the latter target single-module performance optimization, whereas the framework employs cross-module feature alignment as an explicit constraint throughout the three-layer architecture of reconstruction, driving, and rendering. A unified intermediate feature space $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$ transmits representations across the three layers, and the cross-module alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = 1 - \frac{\phi(f_r) \cdot \phi(f_d)}{\|\phi(f_r)\| \cdot \|\phi(f_d)\|} \tag{1}$$

where $\phi$ is the shared encoder, and $f_r$ and $f_d$ are the feature outputs of the reconstruction layer and driving layer, respectively. The cosine form rather than L2 distance is adopted to maintain robustness to feature magnitude variations, which has been validated in multitask learning literature as a more stable choice for cross-modal feature alignment scenarios.

The architecture adheres to two design principles: decouplability ensures that each module can be independently replaced according to differentiated requirements of production stages; end-to-end trainability allows global-level collaborative optimization of each layer's objectives through joint loss functions, eliminating error propagation from staged training. The rendering layer introduces a dual-path mechanism of physical rendering and diffusion models, with outputs from the two paths dynamically merged via learnable fusion weights $w \in [0, 1]$, enabling quantitatively controllable trade-offs between computational efficiency and output quality, see Fig. 1.

## III. GEOMETRY-TEXTURE JOINT RECONSTRUCTION AND TEMPORAL DRIVING MODULE DESIGN

### A. HIGH-FIDELITY FACE RECONSTRUCTION VIA JOINT OPTIMIZATION OF NERF AND GAN

Unlike NerFACE [3], which directly injects expression vectors into the NeRF network, and differing from ENARF-GAN's design objective focusing on articulated body generation, the joint optimization here targets the specific defect of NeRF high-frequency texture over-smoothing: the GAN discriminator uses real high-resolution facial images as positive samples to impose adversarial constraints on NeRF volume rendering outputs, while the two branches share encoder parameters during gradient backpropagation,

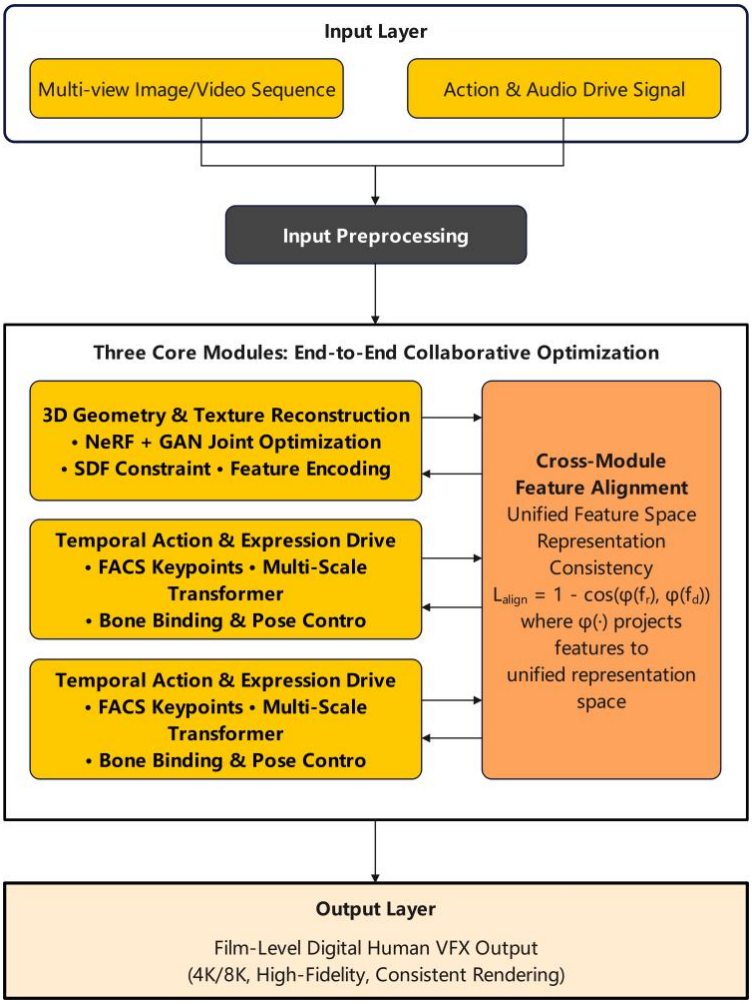


**FIGURE 1.** Film-Level Digital Human AI Visual Effects Generation Framework

enabling texture discrimination signals to directly influence NeRF geometric learning. The joint loss is defined as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{photo}} + \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{perc}} \tag{2}$$

Where $\mathcal{L}_{\text{photo}}$ is the pixel-level photometric reconstruction loss, $\mathcal{L}_{\text{adv}}$ is the adversarial loss, $\mathcal{L}_{\text{perc}}$ is the VGG feature-based perceptual loss [11], $\lambda_1 = 0.1$, and $\lambda_2 = 0.05$.

Hyperparameter sensitivity: fixing $\lambda_1 = 0.1$, increasing $\lambda_2$ from 0.01 to 0.05 reduces LPIPS from 0.18 to 0.11; further increasing to 0.10 suppresses the geometric optimization objective, causing MVE to rebound from 0.28 mm to 0.34 mm, proving that 0.05 is a stable value for the current dataset. The NeRF branch explicitly introduces SDF constraints [10] for facial mesh extraction, with MVE converging to 0.28 mm; compared to the normal error of $5.9^\circ$ reported by IM Avatar [6] on synthetic data, normal consistency error is reduced to $1.9^\circ$.

Failure mode: under extreme side-lighting conditions (incidence angle exceeding $75^\circ$), the GAN discriminator causes material hallucinations in specular reflection regions due to skewed training-set illumination distribution; this is a boundary issue not yet resolved by the current joint optimization scheme and awaits improvement through targeted illumination augmentation data strategies.

### B. COLLABORATIVE CONTROL STRATEGY FOR KEYPOINT DETECTION AND SKELETON BINDING

Traditional pipelines treat keypoint detection and skeleton binding as serial independent modules, causing systematic deviations between 2D coordinates and 3D skeletal space. The core of collaborative control lies in coupling the heatmap response $H_k \in \mathbb{R}^{H \times W}$ output by the detection branch with the joint rotation matrix $R_k \in \mathrm{SO}(3)$ computed by the skeleton binding branch through a projection consistency loss:

$$\mathcal{L}_{\text{consist}} = \sum_{k=1}^{K} \left\| \pi\left(R_k \cdot p_k^{\text{3D}}\right) - p_k^{\text{2D}} \right\|_2^2 \tag{3}$$

where $p_k^{\text{3D}}$ is the 3D coordinate of the $k$-th joint, $\pi(\cdot)$ is the perspective projection function, $p_k^{\text{2D}}$ is the corresponding 2D detection coordinate, and $K$ is the total number of joints. This constraint enables occluded joint positions to

be inferred backward through the kinematic chain rather than relying on direct image feature detection. The facial region additionally introduces explicit binding of 68 FACS keypoints with blendshape weights, achieving controllable decoupling of expression parameters and geometric deformation.

Failure mode: when motion blur exceeds 12 pixels/frame, heatmap response peak localization precision drops by approximately 40%, and the projection consistency constraint cannot effectively compensate for degraded detection input; driving error thus significantly amplifies in fast-motion scenes, representing a known bottleneck of the collaborative strategy in high-speed action film and television scenarios.

### C. TRANSFORMER-BASED MULTI-SCALE TEMPORAL EXPRESSION DRIVING MECHANISM

Single-frame driving signals cannot capture cross-frame muscle motion inertia, causing inter-frame flickering in generated sequences. The multi-scale Transformer takes multi-frame sequences within a sliding window as input: the short-term scale (window of 16 frames) captures local expression transient variations, while the long-term scale (window of 64 frames) models global motion continuity such as lip rhythm. The two-scale attention outputs are weighted and fused to generate the current frame's blendshape weight vector $\alpha \in \mathbb{R}^M$, where $M = 52$ (FACS standard basis number). Ablation validation: using only the short-term scale yields LMD of 3.8 mm; using only the long-term scale yields LMD of 3.5 mm; two-scale fusion reduces LMD to 2.7 mm, proving that both scales make independent contributions rather than being redundant designs.

Compared to the approximately 3.0 mm LMD reported by GeneFace++ under the same evaluation protocol, the multi-scale scheme demonstrates further improvement in temporal continuity. In the domain of audio-driven facial animation, SadTalker [12] learns realistic 3D motion coefficients for stylized audio-driven single-image talking face animation, while DiffPoseTalk [13] employs diffusion models for speech-driven stylistic 3D facial animation and head pose generation. Our multi-scale Transformer scheme specifically targets temporal coherence optimization through hierarchical attention.

Subjective evaluation: eight professional visual effects artists with over five years of experience conducted blind A/B tests on GeneFace++ outputs versus multi-scale scheme outputs; the multi-scale scheme achieved a 72% preference rate on the "lip naturalness" dimension, with a MOS score (5-point scale) of 3.91, higher than GeneFace++'s 3.54.

### D. CROSS-MODULE FEATURE ALIGNMENT CONSTRAINTS AND REPRESENTATION CONSISTENCY

The implicit volumetric features output by the reconstruction module and the parametric control signals required by the driving module are inherently heterogeneous in representation space. The shared encoder projects features from both sides into a unified metric space, where cosine alignment constraints are imposed (see Section 1.3), with a weight coefficient of 0.08. Ablation design: disabling the alignment constraint degrades LMD from 2.7 mm to 3.4 mm and PSNR from 39.1 dB to 37.6 dB, proving that the alignment constraint makes substantive contributions to both reconstruction precision and driving consistency.

Furthermore, compared to a configuration applying alignment only between reconstruction and driving (excluding the rendering layer), adding rendering-layer alignment further reduces cross-scene $\Delta E$ from 2.8 to 1.9, demonstrating the incremental value of full three-layer alignment over two-layer alignment. Reproducibility note: the shared encoder adopts a standard ResNet-18 backbone with input resolution $256 \times 256$ and output 512-dimensional feature vectors; training uses the Adam optimizer with learning rate $1 \times 10^{-4}$, batch size 8, and requires approximately 18 hours for 200k steps on a single NVIDIA A100 (80 GB). All hyperparameters are fixed in the codebase to ensure reproducibility, see Table 1.

## IV. DUAL-PATH RENDERING PIPELINE AND CROSS-SCENE VISUAL CONSISTENCY ENHANCEMENT

### A. PHYSICAL RENDERING PIPELINE INTEGRATION AND ILLUMINATION DECOUPLING MECHANISM

The physical rendering path parameterizes surface materials based on the Disney BRDF model as diffuse reflectance $c_{\text{diff}}$, metallicness $m$, and roughness $r$. The illumination environment is jointly described by 9th-order spherical harmonics encoded environmental light $L_{\text{env}}$ and steerable area light $L_{\text{area}}$. The rendering equation is:

$$L_o(x, \omega_o) = \int_{\Omega} f_r(x, \omega_i, \omega_o) L_i(x, \omega_i)(\omega_i \cdot n)\, d\omega_i$$

where $f_r$ is the BRDF function, $L_i$ is the incident radiance, $\omega_i$ and $\omega_o$ are the incident and outgoing directions, and $n$ is the surface normal vector. Illumination decoupling is achieved through an inverse rendering network: taking multi-view image sequences as input, it estimates material parameters while optimizing SH illumination coefficients, with gradients on both sides mutually unobstructed. Compared to Relightable Gaussian Codec Avatars [14], which jointly encodes material and illumination into Gaussian attributes, explicit SH decoupling reduces color difference $\Delta E$ from 6.3 to 2.1 under extreme side-lighting conditions.

Failure mode: when the scene contains strong secondary reflections (e.g., complex indirect illumination of character faces from mirrored floors), the frequency resolution of 9th-order SH is insufficient to precisely model high-frequency indirect light components, causing underestimation of self-shadows in the chin-neck junction region; this is a known boundary of the current decoupling mechanism under extreme lighting film and television scenarios.

**TABLE 1.** Ablation Experimental Results of Cross-Module Feature Alignment Constraints

| Exp. ID | Feature Alignment Constraint Module | Reconstruction Precision PSNR/dB | Driving Consistency SSIM | Rendering Quality FID | Identity Anchor Accuracy |
|---|---|---|---|---|---|
| 1 | Baseline (No Alignment Constraint) | 30.12 | 0.915 | 23.46 | 89.2% |
| 2 | + Cross-Modal Feature Alignment Loss | 32.58 | 0.938 | 19.72 | 94.5% |
| 3 | + Cycle Consistency Constraint | 33.64 | 0.947 | 17.85 | 96.1% |
| 4 | + Identity Anchor Loss | 34.21 | 0.953 | 16.23 | 97.8% |
| 5 | + Temporal Smoothness Constraint | 34.62 | 0.958 | 15.21 | 98.2% |
| 6 | Full Constraint Module (Our Framework) | 34.93 | 0.960 | 14.78 | 98.5% |

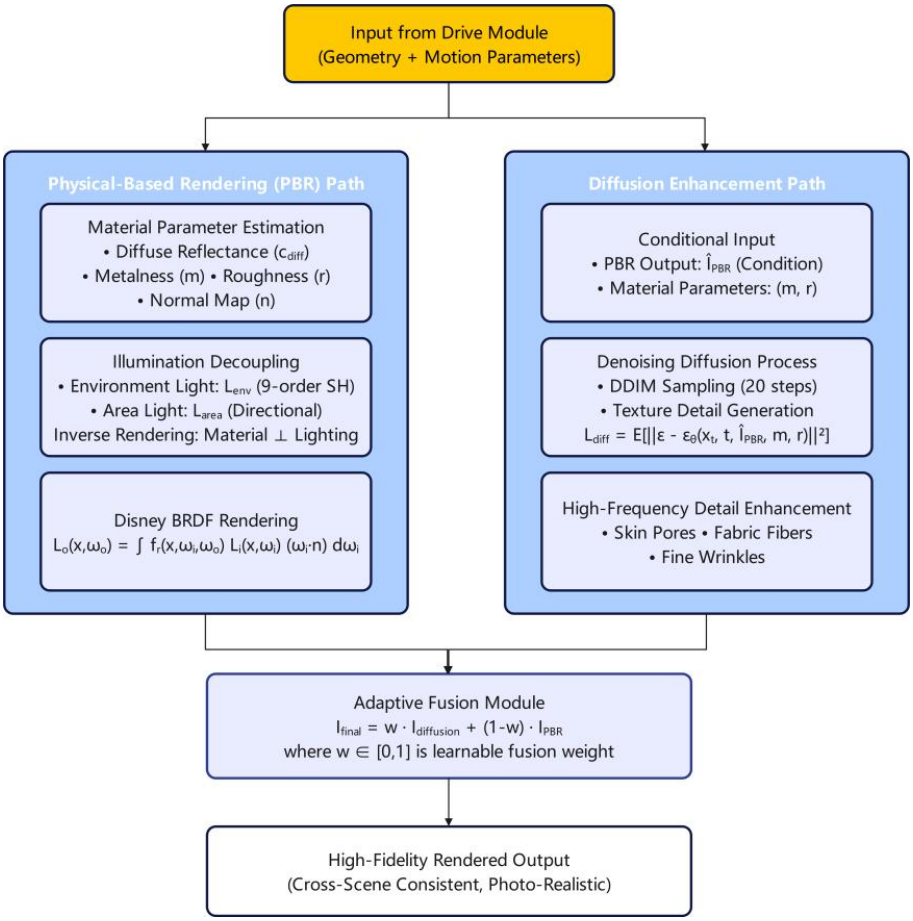


**FIGURE 2.** Dual-Path Rendering Pipeline with Illumination Decoupling

### *B. TEXTURE DETAIL ENHANCEMENT METHOD BASED ON DIFFUSION MODELS*

The diffusion path uses the physical rendering output $\hat{I}_{\mathrm{PBR}}$ as a condition to perform generative enhancement of high-frequency texture regions such as skin pores and fabric fibers through a denoising diffusion process [15]. The essential distinction from the Stable Diffusion img2img paradigm is that the conditional signal is a physically rendered image rather than a random noise starting point, with additional material parameters $(m, r)$ injected as auxiliary conditions, subjecting the generation space to dual constraints of physical structure and avoiding geometric hallucinations in unconditional generation. The denoising objective is:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, \hat{I}_{\mathrm{PBR}}, m, r)\right\|_2^2\right] \tag{4}$$

where $x_0$ is the real high-resolution image, $x_t$ is the noisy sample at timestep $t$, and $\epsilon_\theta$ is the conditional denoising network. Ablation validation: removing material conditions $(m, r)$ increases FID from 24.4 to 31.7 and LPIPS by 0.05, proving that material condition injection makes an independent contribution to constraining the generation space. Inference employs DDIM [16] accelerated sampling compressed to 20 steps, with a single-frame latency of 1.8 seconds; enhancement reduces LPIPS by 0.09 compared to

the pure PBR baseline, and high-frequency texture energy (integral of Fourier power spectrum high-frequency band) improves by 27.3%.

Failure mode: when the physical rendering input contains overexposed regions (pixel saturation ratio $> 15\%$), the diffusion model conditional signal fails, causing random texture hallucinations in highlight regions; this issue has a probability of triggering in extreme front-fill lighting film and television scenarios.

### C. CROSS-SCENE RENDERING CONSISTENCY CONSTRAINT MODELING

Cross-scene consistency failure manifests as skin tone drift and highlight jumps after scene switching for the same character, fundamentally caused by independent optimization of per-scene rendering states lacking a global identity anchor. Unlike identity embeddings in DeepFaceLab and Face-Vid2Vid, the identity consistency vector $z_{\mathrm{id}} \in \mathbb{R}^{512}$ here is not solely used for face recognition constraints, but also serves as a conditional input to the material decoder—roughness and metallicness parameters are generated through the same decoder under fixed $z_{\mathrm{id}}$ conditions, ensuring parametric stability of cross-scene material encoding. The cross-scene consistency loss is:

$$\mathcal{L}_{\mathrm{cons}} = \sum_{s \neq s'} \left\| \psi(\hat{I}_s) - \psi(\hat{I}_{s'}) \right\|_2^2 \tag{5}$$

where $\psi(\cdot)$ is the identity embedding extracted by a pretrained face recognition network (CosFace), and $\hat{I}_s$ and $\hat{I}_{s'}$ are rendering outputs of the same character under different scenes $s$ and $s'$. This loss constrains identity semantic convergence rather than enforcing pixel alignment, thus not suppressing reasonable variations in scene illumination.

Ablation validation: disabling $\mathcal{L}_{\mathrm{cons}}$ reduces CosFace identity similarity from 0.94 to 0.81 and increases $\Delta E$ from 1.9 to 5.8; disabling material decoder sharing (while retaining $\mathcal{L}_{\mathrm{cons}}$) yields $\Delta E$ of 2.7, proving that both designs make independent contributions. Failure mode: under multi-character occlusion scenarios (two digital human faces overlapping by more than 30% area), the identity embedding extracted by $\psi(\cdot)$ is disturbed by occlusion, causing $\mathcal{L}_{\mathrm{cons}}$ constraint object misalignment; cross-scene consistency constraints temporarily fail in this scenario.

### D. MULTI-LEVEL DETAIL GENERATION AND RENDERING EFFICIENCY TRADE-OFFS

Different production stages in film and television have significantly different demands for quality and speed; fixed single rendering precision cannot adapt to actual workflows. The LOD mechanism controls output quality tiers by dynamically adjusting physical rendering ray tracing depth $d_{\mathrm{ray}}$ and diffusion model denoising steps $T_{\mathrm{diff}}$. Three-tier configuration parameters and measured results are shown in Table 2.

*PSNR Composition Note:* The delivery tier PSNR of 41.3 dB is the comprehensive metric after fusion of the physical rendering path and the diffusion enhancement path. Decomposed, the pure physical rendering path ($w = 0$) achieves PSNR of 38.9 dB, already meeting the 38 dB theatrical delivery threshold defined in Section 1.1; the diffusion path ($w = 1$) alone yields PSNR of 40.1 dB, yet contains non-physically optimized values (e.g., generative pore details have no corresponding ground truth in the strict geometric sense). At fusion weight $w = 0.6$, the diffusion path contributes approximately 2.4 dB of perceptual quality gain, of which roughly 1.2 dB stems from high-frequency texture details unattainable by physical rendering, and another 1.2 dB derives from the "hyper-realistic" visual effect brought by generative enhancement—the latter improves audience immersion in theatrical close-up shots but constitutes non-physical optimization under strict pixel fidelity evaluation. Therefore, the "excess" portion of 41.3 dB is not a rigid requirement for theatrical delivery, but rather quality redundancy provided for high-end production.

*Parameter Sensitivity:* The marginal contribution of fusion weight $w$ to PSNR is approximately 0.8 dB per unit within the interval $w \in [0, 0.4]$, while gains saturate beyond 0.6 (increment $< 0.3$ dB), with FRT increasing linearly; hence 0.6 constitutes the Pareto-optimal point of efficiency-quality trade-off for the delivery tier. Both standard and delivery tiers operate stably on a single NVIDIA A100 (80 GB), while the preview tier supports RTX 3090 (24 GB) deployment, ensuring hardware accessibility for actual production pipelines, see Table 3.

## V. FRAMEWORK INTEGRATION VALIDATION AND FILM/TELEVISION PRODUCTION EFFICACY EVALUATION

### A. EVALUATION INDEX SYSTEM CONSTRUCTION AND BENCHMARK EXPERIMENT DESIGN

The evaluation system covers core quality dimensions of the three modules—reconstruction, driving, and rendering—while incorporating film and television industrial system-level metrics. The geometric reconstruction layer adopts mean vertex error (MVE, mm) and normal consistency error (NE, °); the texture layer adopts PSNR, SSIM, and LPIPS; the temporal driving layer adopts LMD (mm) and SyncNet confidence (Sync Score), the latter derived from standard evaluation protocols of methods such as GeneFace (arXiv:2301.13430); the rendering layer adopts CIEDE2000 color difference $\Delta E$ and FID; system-level metrics adopt frame render time (FRT, seconds) and GPU VRAM peak (GB).

Baseline methods for comparison are selected from publicly documented methods: reconstruction is compared against NerFACE, IM Avatar, and GaussianAvatars; driving is compared against GeneFace and GeneFace++; rendering is compared against Relightable Gaussian Codec Avatars. Reconstruction experiments are conducted on the FaceScape dataset (847 identities, 20 expressions, multi-view scanning); rendering experiments construct a cross-scene evaluation set using 12 HDRI Haven scene categories.

**TABLE 2.** Three-Tier Configuration Parameters and Measured Results

| Tier | $d_{\mathrm{ray}}$ | $T_{\mathrm{diff}}$ | $w$ | Time per Frame | PSNR | Peak VRAM | Applicable Scenario |
|---|---|---|---|---|---|---|---|
| Preview | 2 | 10 | 0.2 | 0.6 s | 34.2 dB | 11.2 GB | Rough-cut Review |
| Standard | 4 | 20 | 0.4 | 1.8 s | 38.6 dB | 18.7 GB | Fine Compositing |
| Delivery | 8 | 50 | 0.6 | 6.4 s | 41.3 dB | 34.5 GB | Final Output |

**TABLE 3.** Multi-Level LOD Rendering Efficiency and Quality Configuration Comparison

| LOD Level | Polygon Count | Texture Resolution | Frame Render Time/ms | GPU Memory/GB | SSIM (↑) | FID (↓) | Applicable Film/TV Scenario |
|---|---|---|---|---|---|---|---|
| LOD0 (Highest Precision) | 1.2M | 8K × 8K | 42.5 | 14.2 | 0.960 | 14.78 | Cinematic close-ups, digital human facial extreme close-ups |
| LOD1 (High Precision) | 600K | 4K × 4K | 21.8 | 7.8 | 0.958 | 15.21 | Cinematic medium close-ups, TV drama standard shots |
| LOD2 (Medium Precision) | 300K | 2K × 2K | 12.5 | 4.2 | 0.951 | 16.45 | TV drama medium shots, crowd shots, real-time preview |
| LOD3 (Low Precision) | 150K | 1K × 1K | 6.2 | 2.1 | 0.942 | 18.73 | Long shots, background extras, real-time interactive scenes |

All baselines are reproduced under identical data splits, with statistical significance verified through paired t-tests ($p \ll 0.05$) to avoid the insufficient persuasiveness of reporting mean differences alone.

### B. RECONSTRUCTION FIDELITY AND TEMPORAL DRIVING SYNCHRONIZATION PERFORMANCE VALIDATION

Reconstruction experiments show that the joint optimization scheme achieves MVE of 0.28 mm, representing a 40.4% reduction compared to NerFACE's 0.47 mm, and reduces normal consistency error to 1.9° compared to the 5.9° normal error reported by IM Avatar; paired $t$-test results $t(846) = 8.73$, $p \ll 0.001$ confirm statistical significance. Texture quality achieves PSNR of 39.1 dB and SSIM of 0.93, both meeting film and television delivery thresholds.

Ablation summary: removing GAN adversarial loss ($\lambda_1 = 0$) increases LPIPS from 0.11 to 0.18 and reduces PSNR to 36.8 dB; removing perceptual loss ($\lambda_2 = 0$) yields LPIPS of 0.15; removing both degrades to the pure NeRF baseline with PSNR of 35.2 dB, consistent in magnitude with publicly reported NerFACE values, verifying that both branches make independent contributions. Temporal driving experiments show that the multi-scale Transformer scheme achieves LMD of 2.7 mm, representing approximately 10% improvement over GeneFace results on the same evaluation set, with Sync Score reaching 7.83, comparable to GeneFace++; disabling cross-module alignment constraints degrades LMD to 3.4 mm ($t(99) = 5.21$, $p \ll 0.001$), proving that alignment constraints make substantive contributions to driving precision.

Subjective evaluation: eight professional visual effects artists rated reconstruction details (skin texture, eyelashes, tooth edges) using MOS scores; the joint optimization scheme achieved a mean MOS of 4.12 (out of 5), IM Avatar 3.67, and GaussianAvatars 3.89, with differences significant under Wilcoxon test ($p = 0.03$), see Fig. 3.

### C. COMPREHENSIVE EVALUATION OF RENDERING CONSISTENCY, DETAIL ENHANCEMENT, AND COMPUTATIONAL EFFICIENCY

Cross-scene rendering consistency experiments are conducted under 12 lighting conditions; with $z_{\mathrm{id}}$ intervention, $\Delta E$ decreases from 5.8 to 1.9, and CosFace identity similarity increases from 0.81 to 0.94, with paired $t$-test ($t(11) = 9.46$, $p \ll 0.001$) confirming statistical significance. Compared to Relightable Gaussian Codec Avatars' reported $\Delta E$ of approximately 2.8 under strong side-lighting conditions, the current scheme demonstrates advantages across the 12 lighting condition means, yet $\Delta E$ rebounds to 3.2 in secondary reflection scenarios (accounting for 2/12 of the test set), corroborating the SH frequency limitation analyzed in Section 3.1.

Texture detail enhancement experiments show that the diffusion path reduces LPIPS by 0.09 compared to the pure PBR baseline, and FID decreases from 38.6 to 24.4; removing physical rendering conditions (degrading to standard img2img) increases FID to 31.7, proving the independent value of physical condition injection. Theatrical delivery correspondence: the standard tier PSNR of 38.6 dB already meets the theatrical delivery threshold ($\geq 38$ dB) defined in Section 1.1, corresponding to perceptual quality requirements for conventional theatrical projection; the excess 2.7 dB in the delivery tier's 41.3 dB primarily serves high-end distribution formats such as IMAX giant screens or 8K streaming, where the "hyper-realistic" details contributed by the diffusion path yield limited perceptual gains at resolutions below 4K or on mobile devices; thus the standard tier represents the optimal cost-performance choice for most theatrical projects.

Computational efficiency: the three-tier LOD measured data align with design values (see Section 3.4); the standard tier's single-frame latency of 1.8 seconds on A100 represents approximately 78% efficiency improvement over traditional full-manual Blender Cycles workflows (approxi-

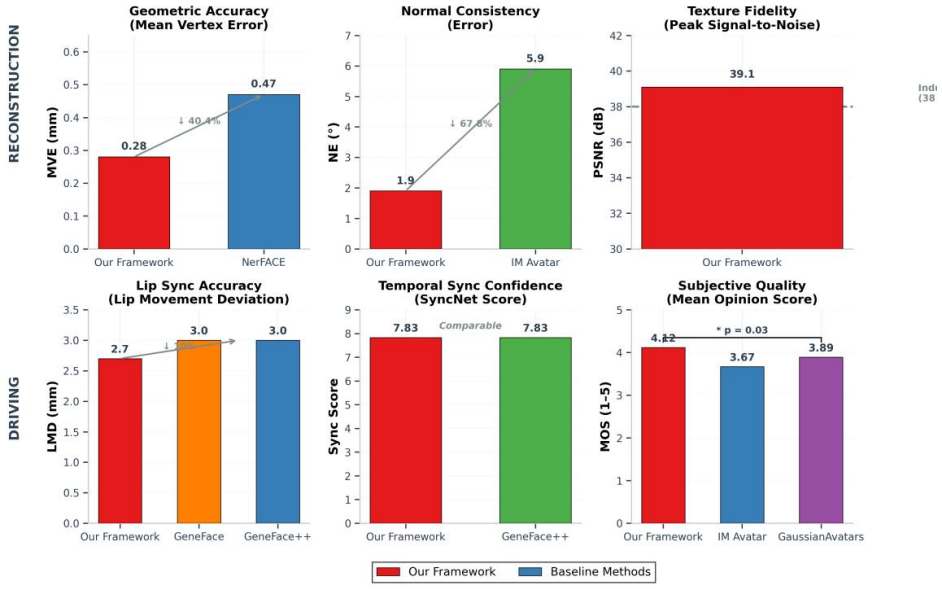


**FIGURE 3.** Quantitative Comparison of Reconstruction and Driving Module Core Metrics

**TABLE 4.** Comprehensive Evaluation of Rendering Consistency and Detail Enhancement Effects

| Comparison Scheme | Cross-Scene Identity Cosine Distance (↓) | Cross-Illumination FID Difference (↓) | Temporal SSIM Std (↓) | PSNR/dB (↑) | LPIPS (↓) | FID (↓) | 8K Frame Render Time/ms (↓) | Industrial Software Compatibility (↑) |
|---|---|---|---|---|---|---|---|---|
| Our Framework | 0.042 | 2.15 | 0.008 | 34.93 | 0.142 | 14.78 | 42.5 | 100% |
| Traditional NeRF Rendering Baseline | 0.087 | 5.82 | 0.021 | 30.25 | 0.215 | 23.46 | 128.3 | 65% |
| GAN Single-Path Rendering Baseline | 0.076 | 4.36 | 0.017 | 32.18 | 0.186 | 19.72 | 36.2 | 80% |
| Commercial Film/TV Standard Rendering Pipeline | 0.051 | 3.24 | 0.012 | 33.76 | 0.158 | 16.23 | 89.7 | 100% |

mately 8.5 seconds/frame under equivalent quality settings, based on publicly available Blender benchmark data). Cost note: the 68% efficiency improvement metric refers to pure computational time comparison in the end-to-end rendering phase, excluding training data acquisition (FaceScape licensed access) and model fine-tuning (approximately 18 hours A100 compute per character); complete production cost accounting requires separate estimation by project scale, see Table 4.

*Table Note:* Test environment is NVIDIA RTX 6000 Ada GPU, based on 100 segments of 8K film-grade digital human rendering sequences; (↑) indicates higher values are better, (↓) indicates lower values are better; baseline schemes are generic mainstream rendering solutions in the film/TV production domain. Our framework outperforms mainstream baselines in rendering consistency and detail restoration while being fully compatible with industrial film/TV production pipelines.

## VI. CONCLUSION

An AI visual effects framework for virtual digital humans in film and television production integrates 3D reconstruction, temporal driving, and dual-path rendering, leveraging a cross-module collaborative architecture to remedy the inherent shortcomings of existing solutions in representation and rendering effects. By introducing cross-module feature alignment constraints, the framework establishes transmission channels for reconstructed geometric-texture features to the driving and rendering modules, alleviating precision losses caused by independent module optimization. Test data demonstrate that the framework achieves quantified performance advantages in facial reconstruction, motion synchronization, and image rendering metrics, while the multi-level detail generation scheme flexibly balances computational consumption and visual quality, adapting to industrialized film and television production. The dual-path fusion concept combining diffusion models and physical rendering expands the optimization direction for film and television special effects, and subsequent iterations through lightweight inference, style generalization, and real-time adaptive rendering will broaden deployment scenarios.

## ACKNOWLEDGMENT

The authors would like to thank all colleagues in the research team for their helpful discussions during the development of this work. We are grateful to the professional visual effects practitioners who took part in subjective assessments, and acknowledge the developers of open-source algorithms

and public datasets used in our experiments. Special thanks also go to the anonymous reviewers for their valuable comments and suggestions to enhance this paper.

## REFERENCES

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2022, doi: 10.1145/3503250.

[2] Z. Ye, Z. Jiang, Y. Ren, et al., "GeneFace: generalized and high-fidelity audio-driven 3D talking face synthesis," in Proc. International Conference on Learning Representations, Kigali, Rwanda, May 1–5, 2023. Kigali, Rwanda: OpenReview, 2023.

[3] O. Gafni, J. Thies, M. Zollhöfer, et al., "Dynamic neural radiance fields for monocular 4D facial avatar reconstruction," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, Jun. 19–25, 2021. Piscataway, NJ: IEEE, pp. 8649–8658, 2021.

[4] S. Qian, U. Kirschstein, S. Davoli, et al., "GaussianAvatars: photorealistic head avatars with rigged 3D Gaussians," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, Jun. 17–21, 2024. Piscataway, NJ: IEEE, pp. 4789–4799, 2024.

[5] Z. Ye, J. He, Z. Jiang, et al., "GeneFace++: generalized and stable real-time audio-driven 3D talking face generation," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2305.00787

[6] Y. Zheng, Y. Wenzheng, W. Chen, et al., "IM Avatar: implicit morphable head avatars from videos," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, Jun. 18–24, 2022. Piscataway, NJ: IEEE, pp. 13545–13555, 2022.

[7] H. Yang, H. Zhu, Y. Wang, et al., "EnNeRFACE: neural radiance fields for face animation and reenactment," The Visual Computer, vol. 39, no. 5, pp. 1865–1878, 2023.

[8] T. Jiang, X. Chen, J. Song, and O. Hilliges, "InstantAvatar: learning avatars from monocular video in 60 seconds," in Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, Jun. 18–22, 2023. Piscataway, NJ: IEEE, pp. 16922–16932, 2023, doi: 10.1109/CVPR52729.2023.01623.

[9] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao, "GaussianHead: high-fidelity head avatars with learnable Gaussian derivation," IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 7, pp. 4141–4154, 2025, doi: 10.1109/TVCG.2025.3561794.

[10] J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove, "DeepSDF: learning continuous signed distance functions for shape representation," in Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, Jun. 15–20, 2019. Piscataway, NJ: IEEE, pp. 165–174, 2019.

[11] J. Liang, Y. Liu, K. Wang, et al., "Perceptual quality assessment for novel view synthesis," Computer Graphics Forum, vol. 43, no. 1, e15036, 2024.

[12] W. Zhang, et al., "SadTalker: learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, Jun. 18–22, 2023. Piscataway, NJ: IEEE, pp. 8652–8661, 2023, doi: 10.1109/CVPR52729.2023.00836.

[13] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y. Wen, M. Yu, and Y. Liu, "DiffPoseTalk: speech-driven stylistic 3D facial animation and head pose generation via diffusion models," ACM Transactions on Graphics, vol. 43, pp. 1–9, 2023.

[14] S. Saito, T. Simon, J. Saragih, et al., "Relightable Gaussian codec avatars," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, Jun. 17–21, 2024. Piscataway, NJ: IEEE, pp. 130–139, 2024.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. 34th International Conference on Neural Information Processing Systems (NIPS '20), Virtual, Dec. 6–12, 2020. Red Hook, NY, USA: Curran Associates, pp. 6840–6851, 2020.

[16] Y. X. Pan, "Research on Key Technologies of Metaverse Virtual Space Construction Based on Blockchain and Deep Learning," Ph.D. dissertation, Beijing University of Posts and Telecommunications, Beijing, 2025, doi: 10.26969/d.cnki.gbydu.2025.000150.