CAG-Avatar：用于高保真头部重建的交叉注意力引导高斯化身

Zhe Chang*、Haodong Jin*、Yan Song*、Hui Yu†
*控制科学与工程系，上海理工大学，中国上海 电子邮箱：
233370890@st.usst.edu.cn,231260086@st.usst.edu.cn, sonya@usst.edu.cn
†心理学与神经科学学院，格拉斯哥大学，英国格拉斯哥 电子邮箱：
hui.yu@glasgow.ac.uk

arXiv:2601.14844v1 [cs.GR] 21 Jan 2026

摘要——**创建高保真、可实时驱动的三维头部化身是数字动画中的核心挑战。尽管三维高斯泼溅（3D-GS）带来了前所未有的渲染速度和质量，但当前的动画技术往往依赖“一刀切”的全局调整方法，即由单一表情代码统一驱动所有高斯基元。这种简单的方法无法揭示不同面部区域各自独特的动态特性，例如可变形皮肤与刚性牙齿之间的差异，从而导致明显的模糊和失真伪影。我们提出条件自适应高斯化身（CAG-Avatar）框架，以解决这一关键局限。其核心是一个基于交叉注意力构建的条件自适应融合模块。该机制使每个三维高斯都能作为查询，根据其规范位置从全局表情编码中自适应提取相关驱动信号。这种“量身定制”的条件化策略显著增强了对细粒度局部动态的建模能力。实验结果证实，该方法显著提升了重建保真度，尤其是在牙齿等具有挑战性的区域，同时保持了实时渲染性能。**

索引词—**3D 高斯泼溅、头部化身、交叉注意力、动态重建、可驱动化身。**

I. 引言

从单目视频中创建高保真、可驱动的数字化身，是实现高级交互系统的基石，应用范围涵盖虚拟现实/增强现实、远程临场，以及智能车辆中的社会视觉 [1], [7], [8]。理想的虚拟形象必须兼顾视觉真实感、采集便捷性和实时性能。然而，由于人头具有复杂的几何结构和动态特性，对其进行重建和动画化仍是一项艰巨挑战。考虑到人类表情的完整范围，这些挑战会进一步加剧：表情不仅包括常见情绪，还包括面瘫等状况下高度复杂、不对称的动作 [2]。这凸显了对能够处理高度局部化和非均匀变形的模型的需求。

以往方法都存在明显的权衡。虽然三维形变模型（3DMM） [3] 能够捕捉全局面部结构，但其线性特性限制了对精细细节和复杂形变的建模，尤其难以表现真实情感表情的细腻之处与微妙差异 [4]。神经辐射场（NeRF） [5]及其变体 [6]–[8]能够实现照片级真实感，但计算开销高得难以承受，因此不适用于实时应用。即使采用显式结构加速的神经辐射场 [10], [11] ，仍受制于体渲染固有的低效率。

近期，3D 高斯泼溅（3D-GS） [12] 通过采用显式的、基于光栅化的表示，彻底革新了实时高质量渲染。这推动了新一波创建可动画头部头像的研究 [16]–[18], [20], [22]。然而，一些方法在动画机制上存在根本性局限，例如 [18]。它们采用“一刀切”的策略，将全局表情编码与每个高斯的特征拼接起来，迫使单个多层感知机学习所有区域特定的响应。这种方法难以解耦面部的混合动态，例如皮肤的非刚性变形和牙齿的刚性运动。

为克服这一瓶颈，我们提出了条件自适应高斯虚拟形象（CAG-Avatar），一种将 3D-GS 的高效性与交叉注意力机制的强大能力相结合的新型框架 [23]。其核心是引入条件自适应融合模块。该模块使每个高斯基元都能充当查询，根据其规范位置，自适应地关注全局表情编码（充当键和值）。这种针对每个基元量身定制的条件化方式显著提升了对细粒度局部动态的建模能力，尤其适用于牙齿等刚性组件。

我们的主要贡献如下：

1) 我们提出了一种基于交叉注意力的融合模块，使每个高斯基元能够根据其位置动态提取相关驱动信号，从而解决全局条件化带来的细节损失和歧义问题。2) 我们设计了一种高效的高保真头部重建框架，将该模块与 3D-GS 相结合，在保持实时渲染性能的同时，显著增强了对混合动态（例如皮肤与牙齿）的建模能力。

II. 相关工作

A. 3D 头部头像重建

3D面部虚拟形象的重建已从传统几何方法发展到现代深度学习方法。一项开创性工作是 3DMM [3],，它将面部形状和外观嵌入低维线性子空间。

尽管许多后续工作 [24], [25], [27] 增强了其表现力，但三维形变模型从根本上受到固定网格拓扑和线性表示的限制，无法捕捉皱纹等高频细节，也无法表现头发等复杂结构。这推动了人们对表现力更强的非线性表示的探索。

B. 基于 *NeRF* 的面部化身

NeRF [5] 的出现通过使用隐式神经函数合成照片级真实感新视角，标志着范式转变。这一强大的能力很快被应用于动态面部化身 [6], [7], [28]。例如，Gafni 等人 [6]提出了一种用于单目四维重建的动态 NeRF。然而，NeRF 的高计算成本源于其密集的光线采样和体渲染机制，因此不适用于实时应用。即使是采用显式数据结构的加速混合方法（如 INSTA [11],），仍受制于光线步进固有的低效率。这推动了人们探索能够完全绕过体渲染的显式表示。

C. 动态化身的三维高斯泼溅

3D-GS [12] 利用各向异性三维高斯的显式表示和高效的光栅化流程，为实时、高质量渲染提供了一条革命性路径。这一突破迅速扩展到动态场景 [13], [15] ，尤其是可驱动头部化身 [18], [30]。这些方法通常通过预测高斯基元从规范空间到目标表情的形变来驱动虚拟形象动画，这一过程通常由 FLAME 等参数化模型引导。

然而，正如我们在引言中所强调的，当前用于头部化身的动态 3D-GS 方法（如 FlashA-vatar [18], ）在变形建模方面存在共同的理论瓶颈。这些方法通常采用直接条件化策略，将全局表情编码与每个高斯的空间位置特征简单拼接。这种空间不变方法对所有基元施加同质的驱动信号，忽略了不同面部区域所需的特定响应。这一简化假设尤其不适用于处理混合动态——即皮肤的非刚性变形与牙齿的刚性运动。它不仅大幅增加了变形网络的学习负担，还会导致口腔区域出现模糊和失真等视觉伪影。因此，如何为每个高斯提供更具判别性的位置自适应驱动信号，仍是本研究旨在解决的核心问题。

III. 方法

A. 预备知识

3D-GS [12] 提出了一种方法，可根据一组具有已知相机参数的图像重建静态场景。场景由一组三维各向异性高斯表示。每个高斯由其中心（均值）$\mu \in \mathbb{R}^3$ 和一个三维协方差矩阵 Σ定义。

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{1}$$

请注意，协方差矩阵 Σ只有在正半定时才具有物理意义，而通过梯度下降进行优化时并不能保证这一性质。为确保可微性并维持这一性质，Kerbl 等人 [12] 将协方差矩阵 Σ 分解为旋转矩阵 R 和缩放矩阵 S。它们分别由一个可学习四元数$\mathbf{r}$（用于旋转）和一个缩放向量 $\mathbf{s}$ （用于缩放）参数化：

$$\Sigma = RSS^\top R^\top \tag{2}$$

给定视图变换矩阵 W 以及投影变换仿射近似的雅可比矩阵 J，一个三维高斯被投影到二维图像平面上。由此得到的二维协方差矩阵 Σ' 按如下方式计算 [19]:

$$\Sigma' = JW\Sigma W^\top J^\top \tag{3}$$

除了几何参数（μ，$\mathbf{r}$，$\mathbf{s}$）外，每个三维高斯还包含两种外观属性：学习得到的不透明度 $\alpha \in [0,1]$ ，以及用于表示其视角相关颜色的球谐函数（SH）系数 $\mathbf{c}$。随后，通过对按深度排序的投影高斯进行阿尔法混合，渲染出给定像素的最终颜色$\mathbf{C}$:

$$\mathbf{C} = \sum_{i \in N} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{4}$$

其中， α_i 是最终计算得到的不透明度。它通过将高斯的学习不透明度与其二维高斯函数在像素中心处的取值相乘得到，该函数使用投影协方差 Σ'计算。

B. 表面嵌入式高斯初始化

以隐式函数为基础的既有头部表示通常采用由 3DMM 驱动的“规范空间加形变”原则。常见的实现方式有两种：直接根据表情编码 [6]对隐式场进行条件化，或将逐顶点形变从 3DMM 的规范空间传播至观测状态 [11], [14]。然而，这种策略存在一个关键局限：未能充分利用 3DMM 网格本身蕴含的丰富几何先验。因此，正如 Xu 等人 [18], 所展示的那样，这类方法难以稳健地建模具有复杂表情的动态化身，因为生成的形变场缺乏必要的结构感知能力。

为克服这一局限，我们将三维高斯绑定到三维形变模型网格表面，确保其随几何结构协调形变。按照 [18],UV 采样进行初始化，构建网格嵌入式高斯场，在保持点布局均匀的同时，对外观和表情驱动形变进行建模。我们的关键创新是通过交叉注意力模块学习一个额外的偏移场，以表示非表面细节；该模块动态调节条件信号，而非拼接全局表情编码。总体框架如图 1 所示。

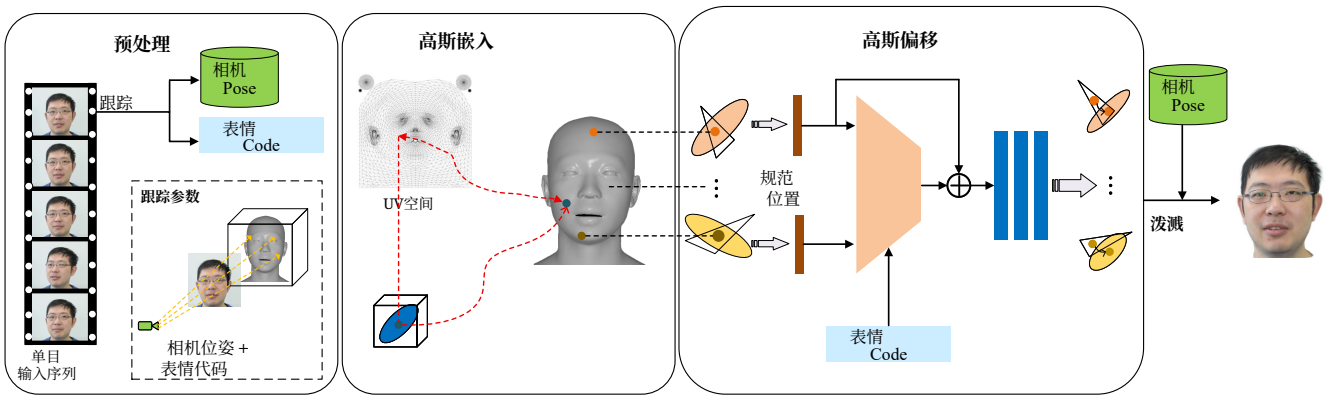


图 1. 所提方法的示意图。我们将三维头部表示为一个高斯场，该高斯场在二维 UV空间中初始化，随后通过网格光栅化嵌入动态 FLAME网格表面。为了驱动高斯点动画，首先将每个高斯中心的规范位置编码为位置特征向量。随后，利用跟踪得到的表情代码，通过交叉注意力融合模块对该特征向量进行细化，其中位置特征充当查询，全局表情编码提供键和值。通过残差连接，将得到的表情条件特征与原始位置特征融合。接着，将这一融合表示输入偏移多层感知机，以预测最终的空间形变，包括位置、旋转和尺度偏移。最后，对形变后的高斯点进行泼溅，在给定的相机位姿下渲染出肖像图像。

如前所述，高斯场可由 $\mathcal{G} = \{\boldsymbol{\mu}, \mathbf{r}, \mathbf{s}, \mathbf{o}, \mathbf{h}\}$ 参数化。通过 UV 采样，我们定义每个网格附着高斯的初始位置，记为 $\boldsymbol{\mu}_M$。在我们的设置中，不透明度 α、球谐系数 $\mathbf{h}$、旋转 $\mathbf{r}$和缩放 $\mathbf{s}$ 均为可学习参数。其中，决定虚拟形象主要外观的前两项属性经过优化后收敛至静态值；几何参数——旋转 $\mathbf{r}$ 和缩放 $\mathbf{s}$——则与额外的位置偏移一同学习，以得到 $\boldsymbol{\mu}_M$。这使模型能够捕捉非表面特征以及面部的精细动态细节。

C. 条件自适应高斯偏移

将高斯基元附着到参数化面部模型上，可以捕捉由表情变化引起的粗粒度几何运动（即从规范位置 $\boldsymbol{\mu}_T$ 到驱动位置 $\boldsymbol{\mu}_M$ 的初始变形），但这不足以表示非表面区域（例如头发）和细粒度的面部动态（例如皱纹和肌肉活动）。更为关键的是，这种粗粒度变形无法区分不同组件的独特运动模式，例如可变形的皮肤与刚性牙齿。

为了解决这一局限，我们不再局限于简单的特征拼接策略，而是设计了一个条件自适应融合模块。该模块的核心是交叉注意力机制，它为每个高斯基元预测动态空间偏移。

具体而言，我们将规范空间中每个高斯基元的位置 $\boldsymbol{\mu}_T$（经过位置编码函数 γ处理后）作为查询。全局表情编码 $\boldsymbol{\psi}$ 同时作为键 和值。通过注意力计算，模型为每个位置生成上下文感知表情特征$\mathbf{c}_\psi$，其中封装了与该基元最相关的驱动信息。

该过程可表示如下：

$$\mathbf{c}_\psi = \text{Attention}(\gamma(\boldsymbol{\mu}_T)W_Q, \boldsymbol{\psi}W_K, \boldsymbol{\psi}W_V) \tag{5}$$

其中，γ 是 Mildenhall 等人的位置编码，[5]、W_Q、W_K、W_V 是可学习投影矩阵。随后，将新生成的上下文特征$\mathbf{c}_\psi$与原始位置编码 γ（$\boldsymbol{\mu}_T$）进行拼接，并输入轻量级多层感知机（MLP） F_θ，以预测位置、旋转和尺度的最终偏移残差：

$$(\Delta\boldsymbol{\mu}_\psi, \Delta\mathbf{r}_\psi, \Delta\mathbf{s}_\psi) = F_\theta(\text{concat}(\mathbf{c}_\psi, \gamma(\boldsymbol{\mu}_T))) \tag{6}$$

最终，目标表情下高斯基元的空间参数由粗略的网格驱动变形与预测的细粒度偏移量之和决定：

$$\mu_\psi, \mathbf{r}_\psi, \mathbf{s}_\psi = (\mu_M \oplus \Delta\mu_\psi, \mathbf{r} \oplus \Delta\mathbf{r}_\psi, \mathbf{s} \oplus \Delta\mathbf{s}_\psi) \tag{7}$$

这种方法取代了传统的“一刀切”条件化策略。例如，对于附着在牙齿区域的高斯基元，我们的模型学会更多地关注表情代码 $\boldsymbol{\psi}$ 中与下颌关节运动相关的部分，同时基本忽略皮肤变形信号，从而保留其刚体运动特征。相反，对于皮肤区域，模型能够精准提取引发皱纹的非刚性变形信号。为每个基元量身定制的驱动信号大幅提升了对混合动态和细粒度细节的建模效果。

D. 训练方案与实现细节

为优化我们提出的框架，我们遵循 [18], 所采用的复合损失函数方案，以确保渲染保真度和训练稳定性。我们的总损失由光度损失 $\mathcal{L}_c$ 和感知损失 $\mathcal{L}_{lpips}$ 组成。

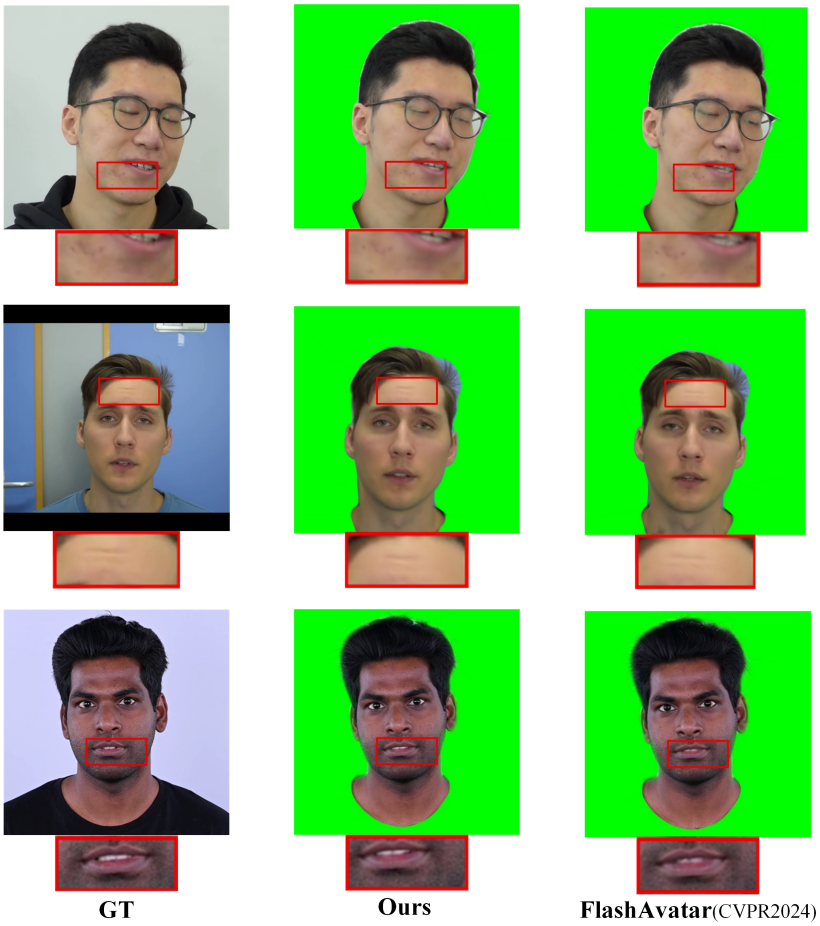


图 2. 与最先进的头部肖像重建方法进行定性比较。我们的模型在重建精细的面部细节和细微表情方面展现出高保真度。值得注意的是，通过所提出的交叉注意力融合机制，与以往方法相比，我们的方法显著改善了对牙齿等具有挑战性的非皮肤区域的重建效果。

损失函数。给定一个表情代码 ψ，我们的模型会生成一套完整的三维高斯参数 $\hat{}\{\mu_\psi, \mathbf{r}^\psi, \mathbf{s}^\psi, o, \mathbf{h}\}$，并据此渲染出一幅图像 I。

为衡量渲染图像 I 与真实图像 I之间的像素级差异，我们采用对异常值具有鲁棒性的Huber损失 $\mathcal{L}_H$（其中 $\delta = 0.1$）[26], 。鉴于口腔区域，尤其是牙齿的重建难度，我们引入口腔掩码 M，并对该区域施加更高的权重 λ_{mouth}，以促使模型捕捉更精细的口腔细节。因此，光度损失 $\mathcal{L}_c$ 定义为：

$$\mathcal{L}_c = \mathcal{L}_H(I, \hat{I}) + \lambda_{mouth}\mathcal{L}_H(I \cdot M, \hat{I} \cdot M) \tag{8}$$

仅依赖逐像素损失可能会导致结果模糊。为提升视觉真实感，我们进一步引入 LPIPS感知损失 [32]，$\mathcal{L}_{lpips}$。该损失利用预训练的 VGG 网络 [33]，在深层特征空间中衡量渲染图像与真实图像之间的相似度。它有助于保留高频细节并稳定训练过程。

最终目标函数 $\mathcal{L}$ 是这两个组成部分的加权和：

$$\mathcal{L} = \mathcal{L}_c + \lambda_{lpips}\mathcal{L}_{lpips} \tag{9}$$

实现细节。我们的框架使用PyTorch [31]实现。我们采用 3D-GS [12]提出的高效可微高斯光栅化器，并利用 PyTorch3D [21] 进行初始的网格到UV空间映射。对于FLAME模型跟踪，我们采用MICA [34],中的分析合成式人脸跟踪器，并结合INSTA [11]的启发进行了进一步改进。表情代码 ψ 由跟踪得到的表情系数、眼部姿态、下颌姿态和眼睑参数拼接而成。

优化设置。 待优化的参数包括三维高斯的颜色（球谐函数）、不透明度、初始旋转和尺度，以及我们提出的条件自适应融合模块和偏移多层感知机中的所有可训练参数。我们使用 Adam优化器 [35] with$\beta = (0.9, 0.999)$。高斯属性的学习率与官方 3D-GS 实现保持一致，而我们新引入的融合模块和偏移多层感知机的学习率设为 $\eta = 1e{-}4$。在训练策略方面，我们

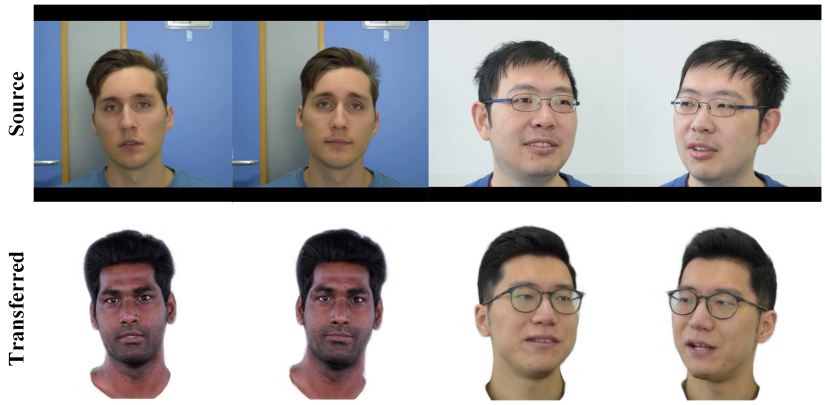


图 3. 我们的方法在人脸重演任务上的定性结果。我们的方法保留了头发、眼睛和嘴部区域中的个性化面部细节，并合成了更加自然的结果。

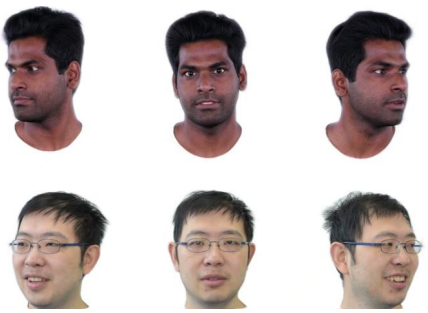


图 4. 新的三维一致性头部姿态合成与表情变化演示。

设置 $\lambda_{mouth} = 40$。为稳定训练初期阶段，感知损失的权重 λ_{lpips} 在前 15,000 次迭代中设为 0，随后增至 0.05。

表 I 与最先进的头部重建方法进行定量比较

方法	L1 ↓	峰值信噪比 ↑	结构相似性指标 ↑	学习感知图像块相似度 ↓
FlashAvatar（CVPR 2024）	0.0137	26.51	0.9256	0.0971
Ours	**0.0128**	**26.53**	**0.9329**	**0.0810**

IV. 实验评估

A. 实验设置

为确保可复现性和公平比较，我们在先前开创性工作使用的公开视频数据集上开展了实验 [6], [11]。所有输入视频序列均经过标准化的预处理流程。首先，将视频帧裁剪并统一调整至 512×512的分辨率。随后，我们采用 Robust Video Matting (RVM) [37] 提取前景主体。为便于准确建模内部结构，我们使用基于 BiSeNet 的解析网络 [36] 对其进行精确分割口腔区域。处理后的视频片段通常时长为一至三分钟。按照标准评估协议，我们将每个序列的最后 500 个视频帧保留为测试集，其余部分用于训练。为确保公平比较，所有实验均在单块 NVIDIA RTX 4090 GPU 上进行。所有方法的训练迭代次数均固定为 15,000 次。

B. 对比分析

我们将本方法与最先进的动态头部虚拟形象框架 FlashAvatar [18],进行基准比较，后者代表了全局条件策略的先前研究成果。如图 2 所示，我们的方法在重建质量方面明显更优。FlashAvatar [18]在牙齿区域存在模糊和几何畸变，而我们的方法能够重建清晰且结构准确的牙齿。此外，对于痤疮印记和皱纹等高频细节，我们的重建结果也更忠实于真实值。表 1 中的定量结果印证了这些定性发现。我们的方法在包括峰值信噪比、结构相似性指标和学习感知图像块相似度在内的所有标准指标上均显著优于基线，为本框架的卓越保真度提供了有力的定量证据。本框架的一项关键优势在于能够对表情和姿态进行精确、解耦的控制。如图 3 所示，在表情重现任务中，我们的模型能够准确再现动态变化，同时保留发型和眼睛等关键身份特征，从而生成自然的结果。此外，图 4 展示了模型在新颖表情与姿态合成方面的泛化能力。

V. 结论

本文解决了基于 3D 高斯泼溅的虚拟形象中的一项关键局限：全局条件化导致局部变形不准确，从而在牙齿等刚性结构上产生伪影。我们提出了一种条件自适应融合模块，利用交叉注意力为每个高斯基元生成独特的空间感知驱动信号。这种设计有效地解耦了表情动态。我们的框架将该模块集成到 3D-GS 中，并取得了

优于基线的结果。实验验证了重建保真度的显著提升，尤其是在牙齿方面，同时保留了实时渲染能力。我们的研究推动了高保真、可驱动数字人的发展，未来工作将聚焦于全身捕捉以及优化注意力机制以提高效率。

致谢

本研究由英国研究与创新署（UKRI）（EP/Z000025/1）和欧洲地平线计划通过玛丽·居里行动（MSCA）资助的 AC-Mod 项目（资助编号：101130271）提供资助。

参考文献

[1] H. Yu, Y. Wang, Y. Tian, H. Zhang, W. Zheng, and F.-Y. Wang, ”面向智能车辆的社会视觉：从计算机视觉到基础视觉,” *IEEE Trans. Intell. Veh.*, vol. 8, no. 11, pp. 4474–4476, 2023年. [2] Y. Xia, C. Nduka, R. Y. Kannan, E. Pescarini, J. E. Berner, and H. Yu, ”AFLFP：一个标注了面部标志点的面瘫数据库,” *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 4, pp. 1975–1985, 2022年. [3] V. Blanz and T. Vetter, ”用于合成三维人脸的可变形模型,” 收录于*Proc. Annu. Conf. Comput. Graph. Interact. Tech. (SIGGRAPH)*, 1999, pp. 187–194. [4] Y. Wang, H. Yu, W. Gao, Y. Xia, and C. Nduka, ”MGEED：多模态真实情绪与表情检测数据库,” *IEEE Trans. Affect. Comput.*, vol. 15, no. 2, pp. 606–619, 2023年. [5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, ”神经辐射场：用于视图合成的场景表示,” *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021年. [6] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, ”用于单目四维面部数字人重建的动态神经辐射场,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8649– 8658.[7] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, ”AD-NeRF：用于说话头合成的音频驱动神经辐射场,” 收录于*Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5784–5794. [8] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, ”HeadNeRF：一种实时的基于神经辐射场的参数化头部模型,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20 374–20384. [9] M. Mihajlovic, A. Bansal, M. Zollhoefer, S. Tang, and S. Saito, ”Key-pointNeRF：通过关键点的相对空间编码泛化基于图像的体积数字人,” 收录于*Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 179–197. [10] Y. Xu, L. Wang, X. Zhao, H. Zhang, and Y. Liu, ”AvatarMAV：利用运动感知神经体素快速重建三维头部化身,” 收录于*Proc. ACM SIGGRAPH*, 2023, pp. 1–10. [11] W. Zielonka, T. Bolkart, and J. Thies, ”即时体积式头部数字人,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 4574–4584.[12] B. Kerbl, G. Kopanas, T. Leimk¨uhler, and G. Drettakis, ”用于实时辐射场渲染的3D 高斯泼溅,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023年. [13] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, ”用于实时动态场景渲染的四维高斯泼溅,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024年, pp. 20 310–20 320.[14] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, ”RigNeRF：完全可控的神经三维肖像,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20 364–20 373. [15] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, ”用于高保真单目动态场景重建的可变形三维高斯,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 20 331–20 341. [16] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang, ”SplattingAvatar：结合网格嵌入式高斯泼溅的逼真实时人类数字人,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 1606–1616. [17] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao, ”GaussianHead：采用可学习高斯扩散的出色头部化身,” *arXiv:2312.01632*, 2023年.

[18] J. Xiang, X. Gao, Y. Guo, and J. Zhang, ”FlashAvatar：采用高效高斯嵌入的高保真头部化身,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 1802–1812. [19] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, ”EWA体积泼溅,” 收录于*Proc. IEEE Vis. (VIS)*, 2001年, pp. 29–38. [20] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu, ”高斯头部化身：通过动态高斯实现超高保真头部化身,” 收录于*Proc. IEEEConf. Comput.Vis. Pattern Recognit. (CVPR)*, 2024年, pp. 1931–1941. [21] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, ”利用PyTorch3D加速3D深度学习,” *arXiv:2007.08501*, 2020年. [22] Z. Zhao, Z. Bao, Q. Li, G. Qiu, and K. Liu, ”PSAvatar：一种利用3D 高斯泼溅实现实时头部化身创建的基于点的可变形形状模型,” *arXiv:2401.12900*, 2024年. [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ”注意力就是一切,” 收录于*Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017年. [24] H. Yu, O. G. B. Garrod, and P. G. Schyns, ”感知驱动的面部表情合成,” *Comput. Graph.*, vol. 36, 162, 2012 3, 152 no. pp. – .[25] Y. Guo, L. Cai, and J. Zhang, ”来自X的3D人脸：从多样化来源学习人脸形状,” *IEEETrans. Image Process.*, vol. 30, pp. 3815–3827, 2021年. [26] P. J. Huber, ”位置参数的稳健估计,” 收录于统计学突破：方法论与分布, 纽约, NY: 施普林格, 1992年, pp. 492–518. [27] L. Zhang, Z. Zhao, X. Cong, Q. Zhang, S. Gu, Y. Gao, R. Zheng, W. Yang, L. Xu, and J. Yu, ”HACK：面向高保真动画的参数化头颈模型学习,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–20, 2023年. [28] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, ”NeRFies：可变形神经辐射场,” 收录于*Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5865–5874. [29] Z. Bai, F. Tan, Z. Huang, K. Sarkar, D. Tang, D. Qiu, A. Meka et al., ”从单目RGB视频中学习个性化高质量体积头部化身,” 收录于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16 890–16 900. [30] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu, ”MonoGaussianAvatar：单目高斯点基头部化身,” 收录于*Proc. ACM SIGGRAPH*, 2024, pp. 1–9. [31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., ”PyTorch：一种命令式风格的高性能深度学习库,” 收录于*Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019年. [32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ”深度特征作为感知度量的非凡有效性,” 收录于 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018 年 , pp. 586–595.[33] K. Simonyan and A. Zisserman, ”用于大规模图像识别的超深卷积网络,” *arXiv:1409.1556*, 2014年. [34] W. Zielonka, T. Bolkart, and J. Thies, ”迈向人脸的度量重建,” 收录于*Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 250–269.[35] D. P. Kingma and J. Ba, ”Adam：一种随机优化方法,” *arXiv:1412.6980*, 2014年. [36] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, ”BiSeNet V2：采用引导聚合的双边网络，实现实时语义分割,” *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021年. [37] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, ”利用时间引导实现稳健的高分辨率视频抠图,” 收录于*Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 238–247.