卷积也需要寄存器：受人类视觉系统启发的动态注意力视频质量评估

Mayesha Maliha R. Mithila 得克萨斯州立大学 圣马科斯，得克萨斯州，美国
mayesha@txstate.edu

Mylene C.Q. Farias 得克萨斯州立大学 圣马科斯，得克萨斯州，美国
mylene@txstate.edu

arXiv:2601.11045v1 [eess.IV] 16 Jan 2026

摘要

无参考视频质量评估（NR-VQA）在没有参考视频的情况下估计感知质量，这往往颇具挑战性。尽管近期技术利用显著性或Transformer注意力，但它们只是将静态图作为辅助输入，而非将上下文从根本上融入视频序列的特征提取中，以此处理视频信号的全局上下文。我们提出了基于全局寄存器的动态注意力视频质量评估（*DAGR-VQA*），这是首个将寄存器令牌 直接融入卷积骨干网络、用于时空动态显著性预测的框架。通过将可学习的寄存器令牌作为全局上下文载体嵌入模型，我们实现了受人类视觉系统启发的动态注意力，生成随时间自适应的显著性图，无需显式估计运动即可追踪随时间变化的显著区域。我们的模型将动态显著性图与RGB输入相结合，以捕捉空间数据，并通过时序Transformer进行分析，从而实现感知一致的视频质量评估。在LSVQ、KonVid-1k、LIVE-VQC和YouTube-UGC数据集上进行的综合测试表明，该模型性能极具竞争力，超过了大多数顶尖基线方法。消融研究表明，融入寄存器令牌有助于构建稳定且在时间上保持一致的注意力机制。在1080p分辨率下达到387.7 FPS的效率，DAGR-VQA展现出适用于多媒体流媒体系统等实时应用的计算性能。

CCS概念

• **计算方法** → **图像与视频采集**；**机器学习**。

关键词

视频质量评估、无参考视频质量评估、动态注意力、寄存器令牌、人类视觉系统

1 引言

随着用户生成视频内容的激增，如何高效、准确地评估视觉质量，已成为流媒体平台、视频压缩系统和内容分发网络等应用的关键。传统的全参考（FR）视频质量评估（VQA）技术依赖于获取未失真参考视频，这在实际部署中往往并不现实。这一局限推动了人们对无参考（NR）视频质量评估方法的兴趣日益增长，此类方法直接从处理后（可能存在失真）的视频序列中估计感知质量 [19, 32, 47]。尽管取得了进展，无参考视频质量评估仍面临重大挑战，因为它必须有效地建模复杂且异质的时空失真，同时与主观质量评分高度一致 [53]。此外，现代流媒体基础设施要求质量评估解决方案具备可扩展性，能够在计算约束下运行，同时在各种内容类型和传输条件下保持准确率 [44]。

人类视觉系统（HVS）是一个复杂的生物系统，由眼部和神经系统组件构成，使我们能够处理视觉刺激。该系统在引导认知注意力关注感知上显著的区域，同时过滤掉无关信息方面发挥着至关重要的作用。视觉显著性预测领域从人类视觉系统中汲取灵感，采用计算模型来识别图像或视频中天然吸引人类注意力的显著区域 [5, 43]。显著区域通常具有明显的运动、强烈的颜色对比度，或包含显著物体。这些模型已成为视频质量评估的重要工具，早期研究通过结合手工设计的时空特征和视觉显著性图，对明显失真进行加权，从而评估质量 [1, 14, 40]。然而，这种方法缺乏近期深度学习（DL）模型所具备的适应性和时序建模能力。

SGDNet [50]采用一种深度学习框架，通过使用预先计算的显著性图进行显著性引导加权，以评估图像质量。该架构侧重于静态显著性，因此不适用于视频应用，因为视觉注意力会随时间动态变化。较新的模型（如 HVS-5M [54], ）则通过采用侧重于内容和边缘依赖性的模块，在特征提取的初始阶段融入视觉显著性。然而，HVS-5M 的静态模块化方法缺乏跨视频序列的动态时间连贯性。相比之下，动态显著性方法会明确建模人类注意焦点如何随时间演变——追踪运动物体、动作和转场，以捕捉观看者的视线 [20, 27]。在视频质量评估中，时间感知至关重要，因为运动和演变事件会不断转移注意力。目前大多数方法只是在后续融合阶段表面地处理动态注意力，而没有将其融入特征提取的核心。

基于Transformer的架构近期取得进展，加之寄存器令牌的应用，展现出有效捕捉视频信号中全局上下文的巨大潜力 [10, 11]。寄存器令牌通过为全局信息提供专用记忆，弥补了标准注意力的一项关键局限。相比之下，分类（CLS）令牌必须同时承担分类和表征任务，而图块令牌则天生具有局部性。对视频质量评估而言，显式建模场景级记忆至关重要，因为一致的全局上下文能够让模型衡量随时间变化的感知质量。然而，在当前的无参考视频质量评估流程中，寄存器令牌通常仅用于Transformer分支，且

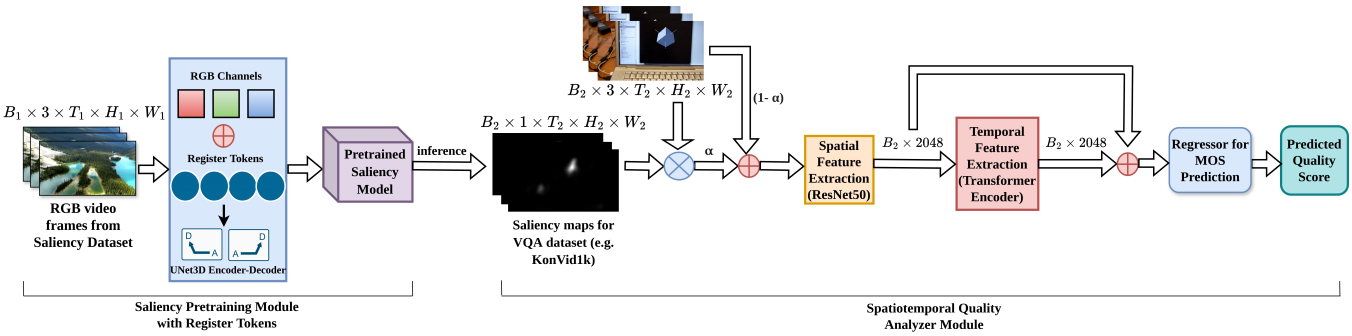


图 1：DAGR-VQA 概览。在预训练期间，将来自显著性数据集（$B_1 \times 3 \times T_1 \times H_1 \times W_1$）的 RGB帧与寄存器令牌拼接（$\oplus$）后输入 UNet3D 进行处理。预训练模型为 VQA视频推理出显著性图（$B_2 \times 1 \times T_2 \times H_2 \times W_2$）。对于每个视频帧，融合操作为 $(1-\alpha)$ 乘以原始帧，再加上（$\oplus$） α 乘以显著性图后与原始帧逐元素相乘（$\otimes$）。空间、时序和回归模块共同完成 VQA流程。

很少用于卷积骨干网络。早期的无参考视频质量评估模型难以捕捉人类注意力如何随内容变化而转移，往往依赖开销高昂的运动分析（例如光流） [5, 58]。后续模型（如 KVQ）将特定区域的动态显著性融入Transformer注意力 [31]。FineVQ则加入了视频和文本的多模态融合 [13]。这些方法在特征提取中缺乏受人类视觉系统启发的动态显著性，因此需要将寄存器令牌等全局上下文先验整合进骨干网络，以连接动态注意力与高效的时空表示，从而实现高性能的视频质量评估。

在这项工作中，我们提出了用于视频质量评估的全局寄存器动态注意力方法（DAGR-VQA）。DAGR-VQA 的整体框架如图 1 所示，由一个寄存器令牌增强的卷积骨干网络（用于提取显著动态特征）、一个空间编码器和一个轻量级时序Transformer组成。我们将寄存器令牌 [10] 作为显式全局先验嵌入卷积 UNet3D 骨干网络中，从而将上下文聚合扩展到视频领域。寄存器令牌有助于整合多个视频帧中的场景级信息，同时保持时间连贯性。这一过程通过在场景推进时将注意力引导至显著区域，模拟人类视觉系统（HVS）。不同于依赖 Transformer 模块 [45, 48] 或显式运动线索 [13, 20], 的既有无参考视频质量评估方法，我们的架构将自适应显著性与轻量级时序建模相融合，从而实现准确的实时预测。主要贡献如下。

- 我们提出一种显著性预训练框架，在卷积特征提取过程中通过寄存器令牌嵌入全局先验。该方法模拟受人类视觉系统启发的动态注意力，无需使用自注意力模块或光流等运动估计方法。
- 我们设计了一种模块化的端到端无参考视频质量评估框架，在骨干网络层面通过寄存器令牌统一全局上下文与质量预测。不同于以往仅将显著性和上下文作为辅助信息的方法，我们的框架直接在骨干网络层面整合二者。
- 我们在四个重要的无参考视频质量评估基准上验证了该方法，结果表现出竞争力。我们还针对寄存器令牌和架构组件进行了消融研究，突显了它们在构建稳定且时序连贯的注意力机制方面的作用。

2 相关工作

深度学习的进展提升了视频质量评估中的时空建模能力，从而提高了预测准确率。然而，由于数据集偏差以及时序建模的局限性，泛化能力仍面临挑战 [53, 56]。近期方法探索了时空注意力 [9] 以及感知线索或显著性线索 [57], , 但这些方法可能计算开销较大且缺乏鲁棒性。为此，研究者对深度特征融合、先进的注意力模块和迁移学习技术进行了研究。例如，ReLaX-VQA [42]采用残差片段和分层特征堆叠来增强时空模式感知，但未显式进行显著性融合。Li 等人 [24] 证明，质量感知预训练能够增强模型对视频失真的鲁棒性，凸显了眼动追踪和感知数据集的价值。

将人类视觉系统原理融入视频质量评估，可确保质量预测符合人类感知，而非仅依赖算法指标。早期方法虽依赖手工设计特征，但深度神经网络已显著提升时空显著性预测的效果。例如，DeepVS [20] 结合了物体与运动线索，而 ViNet-S [16] 则采用由动作信息引导的轻量级 UNet 架构。然而，这类模型仍往往依赖显式运动线索或静态特征，因而难以适应人类注意力的多样化动态变化。鉴于眼动追踪研究表明，剪辑与电影摄影会塑造观众的注意力 [8], , 这一局限尤为值得关注，也促使研究者探索将动态显著性预测作为核心可学习机制加以整合的方法。

预测显著性已越来越多地被用作视频质量评估性能评估的线索 [1, 55]。早期方法如 ST-MAD [40]使用预定义的基于规则的显著性图，依据感知显著性对失真进行加权。大多数早期方法依赖静态显著性，针对每个视频帧计算注意力图，而不进行时序建模；相比之下，动态显著性跟踪跨场景的注意力变化——这对于视频至关重要，因为注意力本质上具有时间依赖性。后来的模型如 SGDNet [50] 和 HVS-5M [54] 纳入了显著性，但仍缺乏完整的时间连贯性。近年来，KVQ [31] 采用融合窗口注意力进行显著性提取，并推进了时序建模。然而，这些方法将显著性视为辅助因素或后期阶段的组成部分，而不是从根本上将其嵌入骨干网络

中。这既反映了架构选择，也反映了大规模数据集的稀缺性；这类数据集包含帧级显著性真值 [2, 8], ，而其缺乏阻碍了骨干网络集成动态显著性的发展。

有效的时空建模对无参考视频质量评估至关重要，因为仅靠静态特征提取无法捕捉随时间演变的感知退化。Freitas 等人 [15] 提出了一种视频质量指标，通过随机森林回归模型整合纹理、显著性、空间活动和时间特征。Vu 等人 [40] 提出了 ST-MAD，这是一种通过强调运动序列中感知上重要区域来捕捉时空失真的模型。该模型依赖手工设计特征，缺乏近期深度学习方法力求实现的精细时空对齐。在此基础上，Ying 等人 [51]提出了 PatchVQ，该方法使用双向长短期记忆网络，从视频块中学习丰富的时空表示。最近，Hu 等人 [18] 通过引入联合时空采样，探讨了视频质量评估中效率与效果之间的权衡，并证明即使对视频进行大幅下采样，轻量级在线视频质量评估仍能给出有竞争力的预测。总体而言，这些工作凸显了时空建模的必要性。

Wu 等人 [44] 提出了 FAST-VQA，并采用网格迷你图块采样（GMS），这是一种自适应机制，能够在保持全局质量评估能力的同时提高计算效率。Morra 等人 [30] 探讨了数据集偏差如何影响用户生成内容的质量评估预测。Sun 等人 [36] 证明，模型复杂度与简洁性之间的权衡取决于数据集构成；自适应特征选择方法能够在保持计算效率的同时增强时空特征提取，而这对于资源受限的视频质量评估系统至关重要。

视觉 Transformer 最初是为图像设计的 [11], ，如今已被应用于视频任务 [3]，并成为视频质量评估中建模时间依赖关系的基础。MS-SCANet [28] 是一种多尺度无参考图像质量评估模型，通过跨分支注意力结合空间与通道特征，但缺乏针对图像的视频帧级时间建模。ViViT [3],等以视频为中心的架构采用时间注意力机制来捕捉运动动态，而 STARVQA [48] 则采用整合式时空注意力方法。自注意力表现出色，但其计算成本会随序列长度呈二次增长，在复杂任务中受到内存限制，且容易受异常影响。DOVER [45],是一种顶尖的基于 Transformer 的无参考视频质量评估模型，采用局部与全局注意力以提升跨数据集泛化能力，但其全局上下文仍局限于注意力层。FineVQ [13] 通过多模态融合和大语言模型提升无参考视频质量评估能力，以进行属性感知评估，并提供可解释质量分数，但弱化了骨干网络层面的动态显著性与全局先验。Darcet 等人 [10] 引入了寄存器令牌，作为可训练记忆单元来保留全局信息。不同于作为分类头的 CLS 令牌和局部图块令牌，寄存器令牌专用于场景级上下文，主要用于大型 Transformer 系统和通用视觉任务，而非轻量级 CNN 骨干网络或感知视频质量评估。

3 方法

如图 1 所示，我们的架构包含两个主要模块：带寄存器标记的显著性预训练和时空视频质量分析器。下文将介绍这两个组成部分。

3.1 基于寄存器令牌的显著性预训练

我们采用 UNet3D 编码器-解码器架构，并引入受视觉 Transformer 启发的寄存器令牌——可学习的全局先验 [10]（图 2）。这些令牌直接集成在卷积骨干网络中，使模型无需自注意力或循环机制即可捕获全局上下文。与 SGDNet [50], 等仅限于静态图像显著性的方法不同，我们的框架明确学习预测时空显著性。相比 ViNet-S [16], ，我们的架构通过广泛使用三维卷积并结合寄存器令牌，更精确地建模运动动态和空间细节。我们完全从头开始在 DHF1K数据集 [41], 上训练显著性网络，确保寄存器令牌能够无缝融入完整的时空处理流程。TASED-Net [26]等现有的预训练方案往往受限于其架构约束，因此不太适用于基于寄存器令牌的模型。尽管 DHF1K 规模适中，但其高质量的眼动追踪标注为监督训练提供了坚实基础。

初始化与拼接。 寄存器令牌作为可学习的全局先验，与输入进行拼接，以引导编码器形成一致的时空模式。设 N 为寄存器令牌的数量， d 为其嵌入维度。寄存器令牌张量$\mathbf{R}$ 初始化如下：

$$\mathbf{R} \sim \mathcal{N}(0,1), \quad \mathbf{R} \in \mathbb{R}^{1\times N\times d\times 1\times 1}, \tag{1}$$

其中，$\mathrm{N}(0, I)$ 表示标准正态分布，该张量以随机方式初始化。随后，使用以下三维卷积将注册令牌投影到 $\mathbf{R}'$ ，以匹配输入的时空维度：

$$\mathbf{R}' = \mathcal{T}(\mathbf{R}) = \mathrm{Conv3D}(\mathbf{R}), \quad \mathbf{R}' \in \mathbb{R}^{N\times T\times H\times W}, \tag{2}$$

其中， T为视频帧的数量， H和W分别为高度和宽度，$\mathbf{T}$ 表示三维卷积变换。

投影后的寄存器令牌与视频输入张量按如下方式进行拼接。增强后的输入 $\mathbf{V}^{\mathrm{aug}}$ 计算如下：

$$\mathbf{V}_{\mathrm{aug}} = \mathrm{Concat}(\mathbf{V}, \mathbf{R}') \in \mathbb{R}^{(C+N)\times T\times H\times W}, \tag{3}$$

其中，$\mathbf{V}$ 是具有 C 个通道的原始视频输入 .

三维U-Net中的卷积滤波器因此能够同时处理局部像素数据和广播的寄存器令牌。寄存器令牌充当紧凑、可学习的记忆，直接将全局先验注入输入中。每个令牌都会被投影为覆盖所有帧的时空映射，使编码器能够在每一层同时评估局部证据和全局上下文。实际上，这些映射充当场景依赖偏置场，引导网络关注最显著且在时间上最稳定的位置。该方法利用常规卷积，在没有显式注意力模块或循环训练的情况下产生类似注意力的效果和全局一致性。

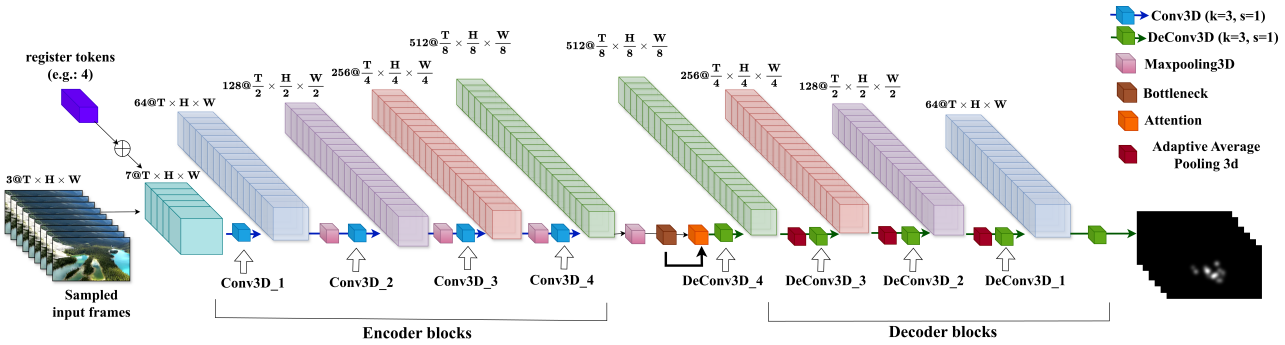


图 2：基于寄存器令牌的显著性预训练架构。采样的输入帧与可学习的寄存器令牌拼接（⊕），并由 UNet3D 式编码器-解码器处理。网络依次应用三维卷积、池化、瓶颈层、注意力和自适应池化，生成时间稳定且具备质量感知能力的显著性图。各阶段的通道数和空间维度标注在模块上方。

编码器-解码器。 在显著性预训练阶段，增强输入序列 $\mathbf{V}_{\text{aug}}$（例如视频与任意辅助通道的拼接）首先由UNet3D编码器处理，记为E。该编码器提取时空特征表示：

$$\mathbf{Z} = \mathcal{E}(\mathbf{V}_{\text{aug}}), \tag{4}$$

其中，E(·) 表示编码层的完整堆栈。

在网络的瓶颈阶段，构建空间注意力图以突出显著区域。具体而言，将编码器输出 Z 输入三维卷积和 Sigmoid激活函数，得到掩码 A(Z)：

$$\mathcal{A}(\mathbf{Z}) = \sigma(\text{Conv3D}(\mathbf{Z})), \tag{5}$$

其中，$\sigma(\cdot)$ 是逐元素 Sigmoid函数。

接下来，瓶颈特征 Z 由变换模块 B（例如三维卷积块）进一步处理，并由注意力掩码调制。Z 是最深编码器层的输出，代表输入的紧凑且高度抽象的版本。因此，精炼后的特征张量 Z′ 为：

$$\mathbf{Z}' = \mathcal{B}(\mathbf{Z}) \cdot \mathcal{A}(\mathbf{Z}), \tag{6}$$

其中，·表示逐元素相乘。最后，解码器 D ˆ 根据经过注意力处理和变换后的瓶颈特征重建预测显著性序列S ：

$$\hat{\mathbf{S}} = \mathcal{D}(\mathbf{Z}'), \tag{7}$$

其中，D(·)表示UNet3D解码器模块。我们的显著性骨干网络将每个视频片段作为三维体积进行分析，并预测其在空间和时间上的显著性。编码器融合各帧的外观与运动线索，并利用共享的寄存器令牌和瓶颈层注意力生成全局先验。解码器对这一表征进行上采样，在一次前向传播中为每帧生成一张显著性图。这便产生了动态显著性，它随运动物体移动，同时保持时间上的连贯性。

损失函数。 为了训练显著性预测模块，我们采用两个互补的损失函数，用于评估预测显著性图 S 与真实显著性图S 在像素级上的匹配程度。令 N 表示每个显著性图中的像素总数。对于索引为 j（$j = 1, 2, \ldots\ldots, N$）的每个像素，$\mathrm{S}_j$ 和 $\hat{}\mathrm{S}_j$ 分别表示真实显著性值和预测显著性值。我们将真实显著性图的均值定义为：

$$\mu_{\mathrm{S}} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{S}_j, \tag{8}$$

以及预测图的均值：

$$\mu_{\hat{\mathrm{S}}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathrm{S}}_j. \tag{9}$$

KL散度损失 [23] 衡量两张图在所有像素上的平均分布差异：

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathrm{S} \parallel \hat{\mathrm{S}}) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{S}_j \log \frac{\mathrm{S}_j}{\hat{\mathrm{S}}_j}. \tag{10}$$

此外，皮尔逊相关系数损失 [4]旨在提升预测图与真实值图之间的线性相关性，其计算方式如下：

$$\mathcal{L}_{\text{CC}} = -\frac{\sum_{j=1}^{N}\left(\hat{\mathrm{S}}_j - \mu_{\hat{\mathrm{S}}}\right)\left(\mathrm{S}_j - \mu_{\mathrm{S}}\right)}{\sqrt{\sum_{j=1}^{N}\left(\hat{\mathrm{S}}_j - \mu_{\hat{\mathrm{S}}}\right)^2}\sqrt{\sum_{j=1}^{N}\left(\mathrm{S}_j - \mu_{\mathrm{S}}\right)^2}}. \tag{11}$$

所有求和与平均运算均针对每张图中由 j 索引的 N 像素进行。

最后，我们将两个目标结合为一个加权损失函数进行训练：

$$\mathcal{L} = \gamma \cdot \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CC}}, \tag{12}$$

其中 γ 是一个超参数，用于控制两个损失项之间的权衡。

3.2 时空视频质量分析器

图 3 展示了时空视频质量分析器模块，该模块整合了显著性驱动的空间编码和基于 Transformer 的时序建模，以生成符合感知特性且时序连贯的视频质量预测。对于输入视频 V，特定时间戳 t 处的视频帧表示为 $\mathbf{I}_t$。我们推导出

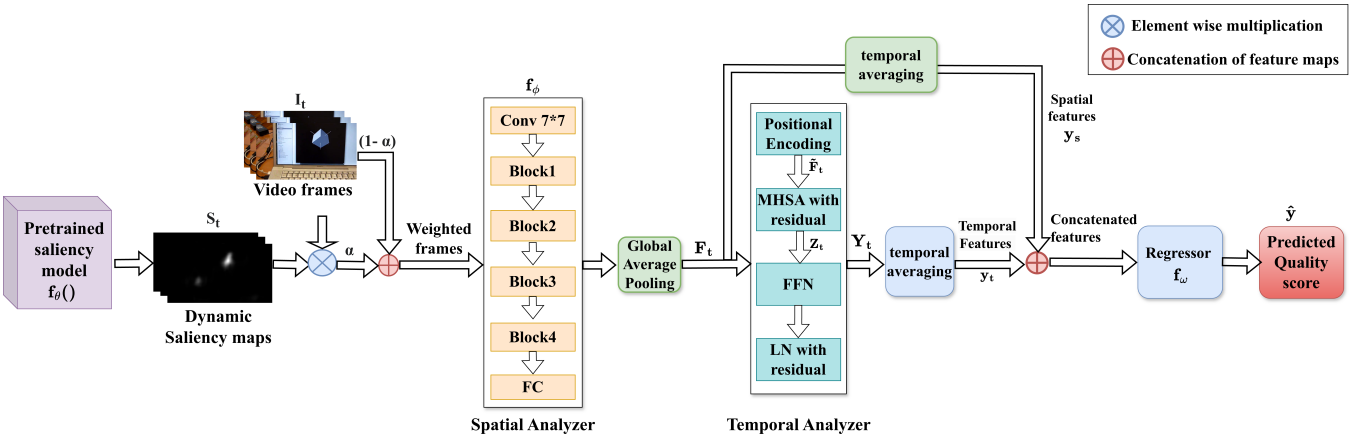


图 3：所提出的时空视频质量评估模块概览。预训练的显著性模型 $f_\theta()$ 会生成动态显著性图，形成显著性加权的视频帧；这些视频帧经过 ResNet-50 骨干网络，以提取并池化空间特征F_t。随后，对这些特征进行时间池化，得到 y_s。接着，序列经 Transformer 编码器处理后，与池化后的时间特征 y_t 拼接并投影，然后输入回归器 f_ω，以预测视频质量分数 $\hat{y}$。LN 表示层归一化，FFN 表示前馈网络，MHSA 表示多头自注意力，Conv 表示初始卷积，Block 1–4 表示 ResNet 卷积阶段，FC 表示全连接层。

使用第 2 节所述的模型，推导出视频帧 $\mathbf{I}_t$的动态显著性图，记为 $\mathbf{S}_t$，如下所示:

$$\mathbf{S}_t = f_\theta(\mathbf{I}_t), \quad \mathbf{S}_t \in \mathbb{R}^{H\times W}, \tag{13}$$

其中， $f\theta(\cdot)$ 是显著性网络。

对于每一帧，我们通过加权求和将 RGB图像 $\mathbf{I}_t$ 与其预测显著性图 $\mathbf{S}_t$ 结合，构造融合输入。随后，使用 ResNet-50编码器从融合输入中提取并池化空间特征 $\mathbf{F}_t$，如下所示。

$$\mathbf{F}_t = \mathrm{GAP}\big(f_\phi\big((1-\alpha)\mathbf{I}_t + \alpha(\mathbf{I}_t \odot \mathbf{S}_t)\big)\big), \tag{14}$$

其中， f_ϕ 表示ResNet-50编码器，GAP(·)表示在空间维度上进行全局平均池化，$\mathbf{I}_t$表示原始帧，$\mathbf{S}_t$表示显著性图， $\odot$表示逐元素乘法， α控制融合。对于每一帧， α设定最终表征中显著性增强信息与原始特征的比例。

接下来，将正弦余弦位置编码（PE）添加到空间特征$\mathbf{F}_t$中，以提供时间顺序，如下所示。

$$\tilde{\mathbf{F}}_t = \mathbf{F}_t + \mathbf{P}_t, \qquad \mathbf{P}_t = \mathrm{PE}(t), \tag{15}$$

其中，$\mathbf{F}_t$是帧 t的融合空间特征，PE(t)表示时间 t的位置编码。为对帧间的时间依赖关系˜建模，我们使用轻量级时间Transformer编码器处理帧 t 的每个位置增强特征向量$\mathbf{F}_t$。该Transformer由两个编码器层组成，每层均包含多头自注意力和前馈网络子模块，以及跳跃（残差）连接和层归一化。

在每个编码器层中，$\mathbf{Z}_t$ 表示帧 t经注意力处理和上下文更新后的表征，并与输入相加形成残差连接，如下所示。

$$\mathbf{Z}_t = \mathrm{MHSA}(\tilde{\mathbf{F}}_t) + \tilde{\mathbf{F}}_t, \tag{16}$$

其中，MHSA(·) 是作用于输入特征 ˜$\mathbf{F}_t$ 的多头自注意力操作。

接下来，该输出$\mathbf{Z}_t$ 经过一个前馈网络处理，并再次加入残差连接，随后进行层归一化，以表示$\mathbf{Y}_t$:

$$\mathbf{Y}_t = \mathrm{LN}(\mathrm{FFN}(\mathbf{Z}_t) + \mathbf{Z}_t), \tag{17}$$

其中，FFN(·) 是前馈神经网络，LN(·) 表示归一化层。

我们首先对每个帧特征图$\mathbf{y}_s$ 沿空间维度进行全局平均池化$\mathbf{F}_t$，然后对这些向量沿时间维度求平均，从而得到平均空间特征。与此同时，我们对 Transformer 输出$\mathbf{y}_t$ $\mathbf{Y}_t$ 沿时间维度求平均，得到平均时间特征。将二者拼接后输入全连接回归层，该层会把此特征表示映射为最终的视频质量预测 $\hat{y}$，如下所示:

$$\hat{y} = f_\omega\big([\mathbf{y}_s; \mathbf{y}_t]\big). \tag{18}$$

视频质量评估训练采用联合损失进行平均意见分数回归，这与显著性预训练损失不同。具体而言，对于平均意见分数预测，我们采用以下加权损失:

$$\mathcal{L} = \mathcal{L}_{\mathrm{L1}} + \beta \cdot \mathcal{L}_{\mathrm{corr}}, \tag{19}$$

其中，L_{L1}是绝对误差损失 $|\hat{y} - y|$，$\mathrm{L}_{\mathrm{corr}}$是斯皮尔曼秩相关损失（$1 - \rho(\hat{y}, y)$）[34],，而 β是固定的损失权重。

这种设置确保感知质量得到一致建模，并能稳健应对数据集中平均意见分数的变化。通过结合显著性增强的空间特征、基于 Transformer 的时间建模和稳健的训练目标，我们的模型能够在不同数据集上生成与感知相符且时间上一致的质量估计。

4 计算复杂度

所提出的 DAGR-VQA 模型旨在通过整合卷积特征提取与时间注意力机制，高效处理视频序列。

为评估其计算成本，我们分析了其核心组件的复杂度，并将其与基于 Transformer 和基于采样的方法进行如下比较。

- 显著性预测（UNet3D）：层次化卷积操作的复杂度为 $O(T\cdot N\cdot d)$，其中 T为帧数，$N = H \times W$ 表示调整大小至 224×398 后每帧的空间令牌数，d 为特征维度。通过单个 Conv3D 操作引入寄存器令牌，额外开销可以忽略不计。
- 空间特征提取（ResNet50）：处理显著性加权帧的复杂度为 $O(T\cdot N\cdot d)$。
- 时序建模（Transformer）：时序Transformer对帧序列引入自注意力，其复杂度为 $O(T^2\cdot d)$。
- 总体复杂度：总计算成本为 $O(T\cdot N\cdot d + T^2\cdot d)$。对于我们的设置（$T = 8$，$N \gg T$），空间处理占主导地位，产生有效复杂度 $O(T\cdot N\cdot d)$。

总之，ViViT 模型在所有时空令牌上采用自注意力机制，因此其复杂度为 $O((T\cdot N)^2\cdot d)$。对于维度为 8×224×224 的输入，该方法需要约 141 GFLOPs。相比之下，DAGR- VQA 模型需要 59 GFLOPs，计算需求降低了 2.4 倍。FAST-VQA [44] 通过采用网格迷你补丁采样实现计算效率，每帧使用 G_f^2 个片段。其 Transformer 在跨越 T 帧进行处理时，主要计算负担的复杂度为 $O(T\times G_f^2\times d)$。在标准配置参数 $G_f = 7$ 和 $T = 8$ 下，计算量为 279 GFLOPs。相比之下，轻量变体 FAST-VQA-M（其特征为 $G_f = 4$）需要 46 GFLOPs，通过大幅下采样实现了计算量降低。相比之下，DAGR-VQA 需要 59 GFLOPs，在计算效率与全面的时空序列处理之间实现了平衡。

5 实验结果

我们在四个标准的无参考视频质量评估数据集上评估了 DAGR-VQA：LSVQ [51], KonVid-1k [17],LIVE-VQC [33],和 YouTube-UGC [37]。这些数据集包含具有不同失真、分辨率和时长的多样化真实世界视频，为评估模型的鲁棒性提供了全面的基准。

我们的实验流程包括显著性预训练，随后进行用于视频质量评估的时空特征提取。显著性模型在 DHF1K眼动追踪数据集上预训练 [41]，共 180 个 epoch，使用 Adam优化器（学习率 5×10^{-3}，批量大小 4，损失权重 $\gamma = 0.01$）。视频质量评估模型使用 Adam（学习率 1×10^{-5}，批量大小 5）和余弦退火调度器训练 300 个 epoch。我们将寄存器令牌的数量设为 $N = 4$，并使用显著性加权因子 $\alpha = 0.5$ 进行空间融合。训练目标结合了 L_1 回归损失和斯皮尔曼相关性，排序损失的权重为 $\beta = 0.1$。两个模块的数据集均按 80:10:10 的比例划分为训练集、验证集和测试集。模型性能通过斯皮尔曼秩相关系数 (SRCC) 和皮尔逊线性相关系数 (PLCC) 进行评估。

为在保持内容保真度的同时优化计算效率，我们遵循近期研究 [18, 36]的建议，采用线性插值进行均匀帧采样。对于显著性预测，每个视频采样 60 帧，以捕捉细致的时空动态。对于视频质量评估，为提高效率，仅选取 8 帧等间距视频帧。采样时，从长度为 L的视频序列中选取N帧并使其均匀间隔分布。具体而言，该算法将视频划分为 N个片段，并从每个片段中选取一个代表帧，以确保整个序列均匀覆盖。所有帧均调整为 224×398 像素。

表 1 总结了我们的方法与四个基准数据集上全面的无参考视频质量评估基线方法的性能。基线方法包括传统的基于特征的方法，以及多种先进的深度学习模型，例如 Patch-VQ [51], FastVQA [44], KVQ [31]。我们还报告了采用领先的预训练显著性模型作为插件式骨干网络的方法变体（例如 UNISAL [12]），这些模型通过时空视频质量评估流程进行集成。我们的方法在所有数据集上始终取得极具竞争力且通常更优的性能。在 LSVQ 上，该方法取得了最高的 SRCC（0.907），超过 FineVQ 和 KVQ，并以 0.892 的 PLCC 排名第三。在 KonVid-1k 上，该方法取得了第二高的 SRCC（0.896），超过 MaxVQA；其 PLCC（0.863）略低于排名前三的方法。在 LIVE-VQC 上，我们的方法取得了最佳 PLCC（0.915）和第二高的 SRCC（0.886），超过 MaxVQA，并紧随 FineVQ 之后。在 YouTube-UGC 上，该方法再次以 0.913 的 PLCC 位居第一，并以 0.910 的 SRCC 排名第二，展现出在用户生成内容上的稳健性能。在所有数据集上取平均后，我们的方法取得了 0.896 的 PLCC 和 0.900 的 SRCC，仅次于 FineVQ，排名第二，并优于 MaxVQA 及其他强基线方法。

这些发现凸显了我们方法的稳健性：它往往优于领先方法，或与之表现相当。在 LIVE-VQC 上，DAGR-VQA 的表现优于 FineVQ（斯皮尔曼等级相关系数为 0.886，对比 0.895）；而在 KonVid-1k 上，FineVQ 的相关性更高（斯皮尔曼等级相关系数为 0.910，对比 0.896）。为严格评估这些差异的显著性，我们针对表2中各数据集的逐视频预测，同时进行了配对t检验和 Wilcoxon 符号秩检验。所有情况下，p 值均远高于标准显著性阈值 0.05，这表明 DAGR-VQA 与 FineVQ 之间观察到的差异在统计学上并不显著。除了预测准确率相当之外，DAGR-VQA 的推理速度更快、吞吐量（帧率）也高于 FineVQ（见表4），因此非常适合实时和大规模场景。

为全面评估我们模型在不同数据集之间的泛化能力，我们进行了5折跨数据集实验，以 LSVQ 作为训练集，并将 KoNViD-1k 和 LIVE-VQC 分别作为不同的测试集。图4展示了各训练/测试配置在五折实验中获得的 SRCC 分布（纵轴），x轴则表示相应的数据集对。每个箱线图展示一种特定方法的表现，中央黑线标示五次运行的 SRCC 中位数。我们将 DAGR-VQA 与四种成熟的无参考视频质量评估基线方法进行了比较：VSFA [25], 基于图像块的PVQ [51], PTM-VQA [52], 和 StarVQA [48]。在两组跨数据库评估（LSVQ→KoNViD- 1k 和 LSVQ→LIVE-VQC）中，我们的模型均取得了最高的中位数

表 1：我们的方法与最先进的无参考视频质量评估模型在四个数据集上的性能比较。每项指标的前三项最佳结果分别以红色、绿色和蓝色标出。

方法	LSVQ 皮尔逊线性相关系数	LSVQ 斯皮尔曼等级相关系数	KonVid-1k 皮尔逊线性相关系数	KonVid-1k 斯皮尔曼等级相关系数	LIVE-VQC 皮尔逊线性相关系数	LIVE-VQC 斯皮尔曼等级相关系数	YouTube-UGC 皮尔逊线性相关系数	YouTube-UGC 斯皮尔曼等级相关系数	平均值 皮尔逊线性相关系数	平均值 斯皮尔曼等级相关系数
BRISQUE [29]	0.576	0.579	0.658	0.657	0.638	0.593	0.395	0.382	0.567	0.553
VIDEVAL [38]	0.783	0.794	0.780	0.783	0.751	0.752	0.773	0.779	0.772	0.777
StarVQA [48]	0.857	0.851	0.796	0.812	0.808	0.732	N/A	N/A	0.820	0.798
Patch-VQ [51]	0.828	0.827	0.786	0.791	0.837	0.827	N/A	N/A	0.817	0.815
FastVQA [44]	0.877	0.876	0.855	0.859	0.844	0.823	0.852	0.855	0.857	0.853
MaxVQA [46]	N/A	N/A	0.895	0.894	0.873	0.854	0.890	0.894	0.886	0.881
PTM-VQA [52]	0.864	0.855	0.872	0.857	0.820	0.811	0.857	0.858	0.853	0.845
DOVER [45]	0.889	0.888	0.906	0.909	0.875	0.860	0.891	0.890	0.890	0.888
FineVQ [13]	0.900	0.900	0.915	0.910	0.895	0.895	0.910	0.914	0.905	0.905
KVQ [31]	0.897	0.896	0.892	0.890	0.843	0.820	N/A	N/A	0.877	0.868
UNISAL [12] + 时空视频问答	0.839	0.872	0.889	0.859	0.891	0.872	0.876	0.875	0.874	0.870
DiffSal [49] + 时空视频问答	0.847	0.879	0.885	0.867	0.893	0.874	0.883	0.884	0.877	0.876
ViNet-S [16] +时空视频问答	0.850	0.885	0.887	0.872	0.897	0.883	0.890	0.887	0.881	0.882
Ours	0.892	0.907	0.863	0.896	0.915	0.886	0.913	0.910	0.896	0.900

表2：比较 DAGR-VQA 和 FineVQ 的 p值。

Test	LIVE-VQC	KonVid-1k
配对t检验	0.41	0.094
Wilcoxon符号秩检验	0.39	0.088

和总体 SRCC，优于所有其他方法。这表明，DAGR-VQA 不仅在各个数据集上表现优异，而且对新的视频内容和以往未曾遇到的失真类型也具有稳健的泛化能力。

为评估在我们的显著性框架中采用寄存器令牌增强的效果，我们在 DHF1K 和 DIEM 两个验证集上对模型进行了基准测试，采用了三种成熟的显著性指标：归一化扫描路径显著性（NSS） [21], , 用于量化预测显著性图与人类注视位置之间的对应程度；相关系数（CC） [7], , 用于评估预测图与真实值图之间的线性关系；以及 Judd 曲线下面积（AUC-J） [6], , 用于衡量预测注视点的 ROC 曲线下面积。如表 3 所示，我们完整的 UNet3D+RT 模型显著优于近期的先进方法，包括 ViNet-S、TASED-Net，以及我们自身未采用寄存器令牌的基线。在 DHF1K 上，我们的模型取得了 3.683 的 NSS、0.704 的 CC 和 0.942 的 AUC-J；在 DIEM 上，相应数值为 2.856、0.685 和 0.902，在两个数据集上均创下了新的性能标准。值得注意的是，相较于未采用 RT 的 UNet3D 及其他竞争方法，我们的模型在所有三项指标上均稳定提升，凸显了加入全局寄存器令牌所带来的可靠收益。这些结果证实，加入寄存器令牌能够使网络整合全场景范围内的长距离信息，从而在动态场景中生成与人类注视点显著更加吻合的显著性预测。

我们在十个1080p视频（1920x1080）上，将 DAGR-VQA 的计算效率与当前最先进的 VQA 方法进行了比较，每个视频包含 240 帧。为确保与学界的报告标准一致

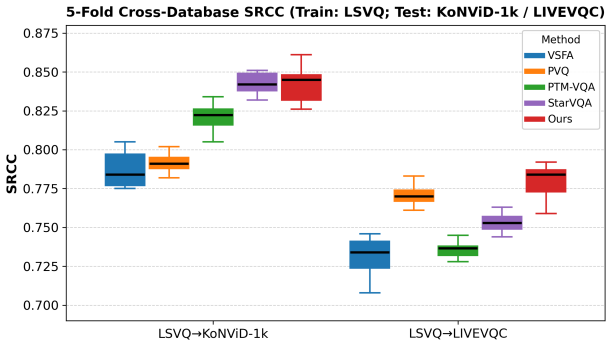


图4：五折跨数据库 SRCC（训练：LSVQ → 测试：KoNViD-1k / LIVEVQC）。每个箱线图展示一种视频质量评估方法的结果，其中箱体内的黑色中心线表示 5 次运行结果的 SRCC 中位数。

, [13, 44, 45],每秒帧数（FPS）值均根据测试帧总数（具体而言为 $10 \times 240 = 2400$）进行归一化，而不考虑实际的帧采样方式。推理时间直接在启用混合精度的 RTX A5000 GPU 上测量。如表4所示，DAGR-VQA 在 1080p 下实现了 387.7 的归一化 FPS，与近期的 VQA 基准相比具有较强竞争力。所有 FPS 值均按照可用视频帧总数进行归一化，以便与先前工作 [13, 44, 45] 直接比较。除具备较高的系统级效率（见表4）外，DAGR-VQA 在四个主要的无参考视频质量评估基准上还取得了总体第二高的准确率，平均 SRCC 为 **0.900**，PLCC 为 **0.896**（见表 1）。

6 消融研究

为了更好地理解动态显著性建模和寄存器令牌对视频质量评估准确率的单独及联合贡献，我们在四个基准

表 3：在 DHF1K 和 DIEM 验证集上对显著性预测模型进行的定量评估。最佳结果以粗体标出。

指标	ACLNet		ViNet-S		TASED-Net		DeepVS		DiffSal		UNet3D（本文方法，不含 RT）		**UNet3D+RT（本文方法）**	
	DHF1K	DIEM	DHF1K	DIEM	DHF1K	DIEM	DHF1K	DIEM	DHF1K	DIEM	DHF1K	DIEM	DHF1K	DIEM
NSS	2.35	2.02	3.008	2.732	2.706	2.16	2.774	2.25	3.066	2.65	2.945	2.720	**3.683**	**2.856**
CC	0.434	0.522	0.529	0.673	0.481	0.557	0.525	0.49	0.533	0.66	0.640	0.675	**0.704**	**0.685**
AUC-J	0.890	0.869	0.919	**0.908**	0.894	0.881	0.887	0.86	0.918	0.907	0.904	0.897	**0.942**	0.902

表4：10 个视频（每个视频包含240帧）在 1080p 下的推理时间和帧率。最佳结果以粗体标出。

指标	TLVQM[22]	VIDEVAL[38]	VSFA[25]	SimpleVQA[35]	FastVQA[44]	Dover[45]	FineVQ[13]	DAGR-VQA
推理时间（秒）	476.5	1059.2	53.6	40.3	**7.39**	**5.07**	10.46	**6.19**
FPS	5.04	2.27	44.80	59.6	**324.7**	**473.4**	229.4	**387.7**

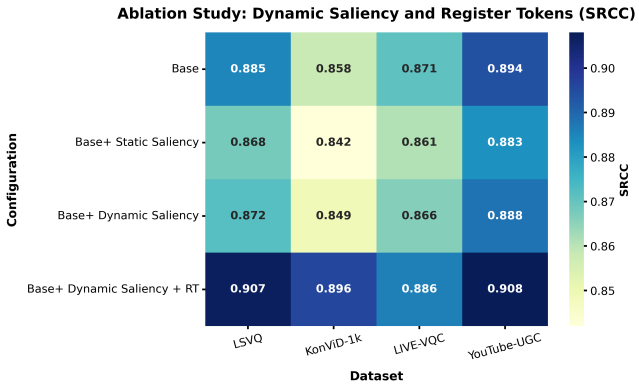


图5：消融研究：在四个数据集上，采用/不采用动态/静态显著性和寄存器令牌的各配置的斯皮尔曼等级相关系数。数值越高越好。

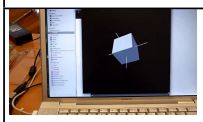
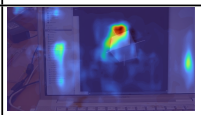
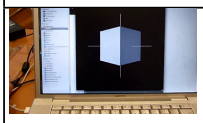
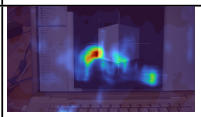
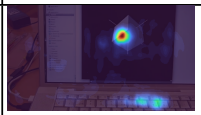
数据集上进行了消融研究。图5通过热图直观总结了所有主要模型变体的斯皮尔曼等级相关系数值。结果表明，基础模型提供了较强的基线性能，而单独添加静态或动态显著性所带来的准确率相近或略低，这可能是因为在某些视频序列中注意力聚焦不够完善。然而，将寄存器令牌与动态显著性结合使用后，所有数据集上的斯皮尔曼等级相关系数均持续显著提升，在 LSVQ（0.907）、KonVid-1k（0.896）、LIVE-VQC（0.886）和 YouTube-UGC（0.908）上均取得最高性能。这些发现凸显了寄存器令牌能够稳健地提升动态显著性整合的稳定性和有效性，使模型能够更好地关注与内容相关的区域，并保持强大的泛化能力。

随后，表5对所提出模型为 LSVQ 和 KonVid-1k 数据集中特定视频帧生成的显著性图像进行了定性分析。这些图示展示了采用寄存器令牌的动态显著性预测如何保留时间连贯、且能适应不同场景语境的注意力。例如，在一段来自 LSVQ 数据集的视频序列中，图像显示模型能够可靠地追踪篮球比赛中始终最显著的运动，从持球者转向“球场上的动态”，最终聚焦于篮筐处最显著的时刻，展现出流畅且实质性的显著性转变。在 KonVid-1k 示例中，寄存器令牌有效地引导模型的注意力遍及旋转立方体，突出边缘、顶点和受光表面等高对比度且与运动相关的区域，同时也纳入次要的视觉干扰。与不使用寄存器令牌的基线相比，这些视觉 发现表明，空间稳定性和时间一致性均有所提升。因此，动态显著性预测能够无缝适应物体运动和场景语义，贴近人类视觉注意力。

为了评估寄存器令牌所封装的全局内容的敏感性，我们从训练好的模型中提取了 DHF1K 数据集中每个视频对应的全部 4 个寄存器令牌的嵌入。为封装针对每个视频学到的全局上下文，我们计算与该视频相关的标记嵌入的均值，从而为每个视频得到一个池化向量。接下来，我们采用 t-SNE [39], ，这是一种非线性降维方法，可将高维特征映射到二维空间，同时确保保留数据点之间的局部相似性。图6展示了 DHF1K 中每个视频的寄存器令牌嵌入均值池化后的二维 t-SNE 投影。图中的每个点代表一个视频，其颜色反映了 DHF1K 元数据所标示的内容类型。该图显示，来自同一内容类别的视频会聚集在一起，尽管训练过程中并未使用类别信息。这种聚类表明，学到的寄存器令牌嵌入编码了相关的内容信息，使显著性预测模块中的注意力更具一致性和内容感知能力。

接下来，我们考察寄存器令牌数量（记为 N）对显著性预测准确率以及下游视频质量评估任务性能的影响。在视频质量评估分析中，我们在基于 DHF1K 预训练的显著性骨干网络中改变 $N \in \{2, 4, 8, 16\}$，随后评估 DAGR-VQA 在 LSVQ、KonVid-1k、LIVE-VQC 和 YouTube-UGC 数据集上的结果。图 7 展示了 SRCC 值随 $\log_2(N)$ 变化的曲线。SRCC 从 $N = 2$ 提升至 4，但在 $N \geq 8$ 时下降，这表明适量的令牌数量更有利，而令牌数量过多则会增加冗余并降低表征区分度。我们观察到，当

表5：LSVQ 和 KonViD-1k 样本的动态显著性与寄存器令牌。每组三张图分别展示原始帧、叠加动态显著性热图（含 RT）后的图像，以及简短的叙述。模型的注意力会随时间自适应转移，跟随动作和场景上下文，体现出真正受人类视觉系统启发的时间反应性。

LSVQ：借助寄存器令牌实现动态显著性		
原始帧	**显著性叠加图**	**叙述**
		显著性聚焦于“发起进攻的白色球衣球员”以及随着球队推进，“持球者”上面的（区域）[VALUE]后场。
		注意力跟随“持球者穿过中场”，灵活追踪进攻动作，并适应球员和球的位置变化。
		在此期间，显著性集中于“黑色球衣球员”比赛高潮，突出防守动作以及他们站位的重要性。
KonViD-1k：基于寄存器令牌的动态显著性		
原始帧	**显著性叠加图**	**叙述**
		立方体旋转时，注意力会被其明亮的顶部边缘吸引，该边缘是视觉上最突出的运动部件。
		随着立方体转动，显著性向左移动，以跟随新显现的面，并适应物体运动和方向变化。
		焦点集中在立方体的正面，随着它转向背面，模型会随着每次变化调整，关注视觉上最突出的表面。

评估基于显著性骨干网络本身的视频质量评估性能。在 DHF1K 上，视频质量评估性能在 $N = 4$ 时达到峰值，并在$N \geq 8$时下降；而在 DIEM 上，最佳视频质量评估性能出现在$N = 2$，$N = 4$ 仍具有竞争力，但更高的值又会导致准确率下降。这支持了这样一种直觉：N控制着全局先验的容量：较小的 N 有助于形成紧凑且正则化程度较高的表征（这更适合规模较小、类别多样性较低的 DIEM）；而略大的 N则有利于多样性更高的 DHF1K数据集以及复杂的真实世界视频质量评估基准。总体而言，这些结果表明，最佳寄存器令牌数量处于较低范围（$N \in \{2, 4\}$），能够在两个显著性数据集和全部 4 个视频质量评估数据集上实现良好的权衡。

我们通过在图8中引入显著性控制参数 α，探究基于显著性的加权对定量视频质量评估性能的影响。 α的取值范围从 $\alpha = 0$ 到 $\alpha = 1$，其中$\alpha = 0$ 表示仅使用原始特征，$\alpha = 1$ 表示仅使用显著性增强特征。图8展示了显著性参数α（位于x 轴）与斯皮尔曼等级相关系数（位于y 轴）得分之间的关系，体现了模型在不同 α取值下的预测准确率。结果表明，将 α从 0.0（不使用显著性权重）提高到 0.5，可使所有数据集上的斯皮尔曼等级相关系数持续提升，并在 $\alpha = 0.5$时达到最大值。然而，对于 $\alpha > 0.5$，性能持续下降，这表明显著性权重过高可能会抑制来自不太显著区域的相关线索。这些发现证实，均衡地分配权重，特别是在 $\alpha = 0.5$时，有助于模型关注面部和运动物体等具有重要感知意义的区域，同时利用有价值的全局信息。

最后，为了更好地了解各个组件（空间分析、时序建模和显著性融合）如何为最终的视频质量模型作出贡献，我们在 LSVQ 和 KonVid-1k 数据集上进行了消融研究。研究结果见表6，表明单独进行空间建模或时序建模都不足以满足需求。仅采用空间建模或仅采用时序建模的配置得分明显较低，尤其是在 KonVid-1k 上。将寄存器令牌增强的动态显著性图加入空间特征提取以提供引导后，性能有所提升，凸显了关注每个视频帧中重要区域的价值。然而，将三个要素（空间、时序和显著性）结合起来，准确率便大幅提升，达到最佳水平；完整模型在 LSVQ 和 KonVid-1k 上的 SRCC 分别为 0.892 和 0.863，且在两个数据集上均取得最高的 PLCC 分数。这一递进过程表明

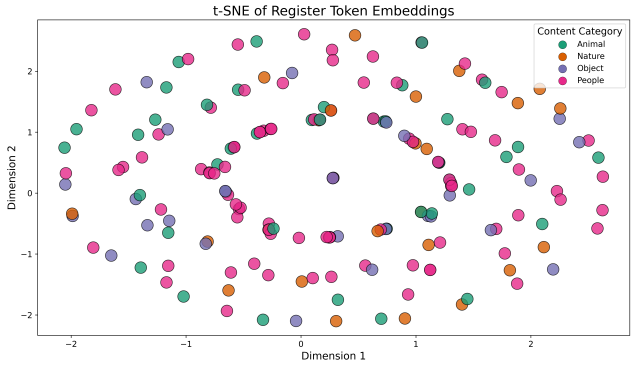


图6：寄存器令牌嵌入的 t-SNE 图，尽管没有语义监督，但会根据内容进行软分组。

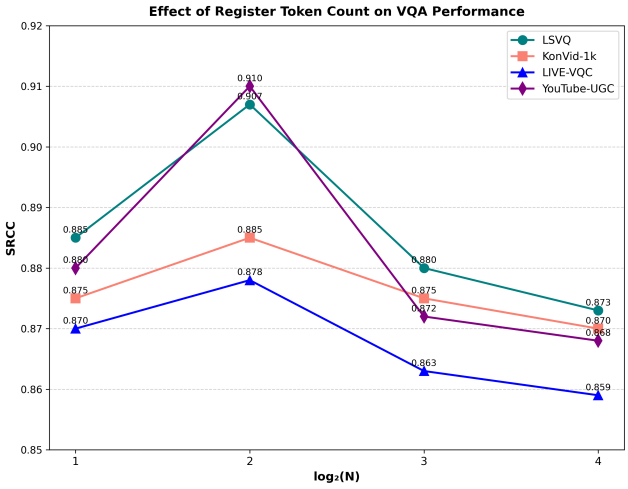


图7：在四个数据集上，寄存器令牌数量对斯皮尔曼等级相关系数的影响，x 轴为 $\log_2(N)$，y 轴为斯皮尔曼等级相关系数（DHF1K）。所有情况下，斯皮尔曼等级相关系数均在 $N = 4$ 处达到峰值，表明适量的寄存器令牌可带来最佳的视频质量评估性能。

每个模块都针对视频质量的独特方面发挥作用，整合全部三个模块对于获得最可靠的评估结果至关重要。

局限性与失效案例。我们的方法在主体运动明显或交互程度适中的场景中表现良好。不过，它也存在一些性能局限。在大多静止的场景或空间位移较小的场景中，投影后的显著性图往往仍集中于少数区域，在视频帧之间几乎没有动态变化，因此相比更简单的静态方法优势不大。相反，在有多个运动主体或活动较为复杂的场景中，模型偶尔会生成破碎或分散的显著性区域，同时突出多个位置，或在视频帧之间出现轻微的不一致。我们的消融研究还表明，显著性图的质量和模型的整体性能会受到注册令牌数量 N和显著性控制参数 α的影响。谨慎

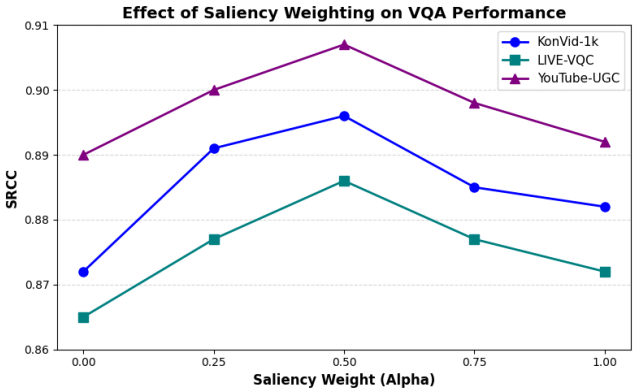


图8：显著性加权（α）对三个数据集上斯皮尔曼等级相关系数的影响，x 轴为加权系数 α，y 轴为斯皮尔曼等级相关系数。性能在 $\alpha = 0.5$ 处达到峰值，表明适度的显著性可带来最佳的视频质量评估结果。

选择或调整这些参数，或许有助于最大限度地减少上述局限。

表6：在 LSVQ 和 KonViD-1k 验证集上评估的时空视频质量评估分析器组件消融研究。每种配置均单独考察或组合关键模块，以展示它们对视频质量预测的独立贡献和整体贡献。

配置	LSVQ		KonViD-1k	
	斯皮尔曼等级相关系数	皮尔逊线性相关系数	斯皮尔曼等级相关系数	皮尔逊线性相关系数
仅空间	0.825	0.832	0.813	0.826
仅时间	0.872	0.877	0.838	0.870
空间 + 显著性融合	0.865	0.871	0.820	0.866
空间 + 时间 + 显著性	**0.892**	**0.907**	**0.863**	**0.896**

7 结论

我们提出了 DAGR-VQA，这是一种新颖的无参考视频质量评估框架，将寄存器令牌作为全局先验，并结合卷积显著性骨干网络，以有效预测动态显著性。结果表明，这一策略能够改进时间自适应显著性建模，并增强模型识别各类视频内容中感知退化的能力。在多个公开视频质量数据集上的实验表明，DAGR-VQA 的准确率表现具有竞争力，且泛化能力良好。消融研究显示，对寄存器令牌进行经验性调优，或可使模型更可靠地关注视频序列中最相关的区域和时刻。DAGR-VQA 的高效性和模块化结构有助于将其快速集成到流媒体系统框架中，实现实时视频质量监测。未来工作可探索自适应显著性融合方法及更广泛的现实应用。

参考文献

[1] Welington YL Akamine 和 Mylène CQ Farias。2014。使用视觉注意力计算模型进行视频质量评估。《电子成像期刊》23，6（2014），061107–061107。[2] Hani Alers、Judith A Redi 和 Ingrid Heynderickx。2012。利用显著性图考察任务对视频观看行为的影响。载于《人类视觉与电子成像 XVII》，第 8291 卷。SPIE，269–276。[3] A. Arnab、T. Han 和 A. Vedaldi。2021。ViViT：视频视觉Transformer。载于IEEE计算机视觉与模式识别会议（CVPR）论文集。IEEE。https://doi.org/ 10 1109 2021 3074933 。/TPAMI.。Jacob [4] Benesty、Jingdong Chen、Yiteng Huang 和 Israel Cohen。2009。皮尔逊相关系数。施普林格。1–4 页。https://doi.org/10.1007/978-3-642- 00296-0 5[5] Ali Borji 和 Laurent Itti。2012。视觉注意力建模的最新进展。《IEEE模式分析与机器智能汇刊》35，1（2012），185–207。[6] Ali Borji、Hamid R. Tavakoli、Laurent Itti 和 Christof Koch。2013。视觉显著性预测中的评分、数据集和模型分析。载于 ICCV。[7] Neil D.B. Bruce、Christian Wloka、Nicholas Frosst 和 Sajib A. S. Rahman。2015。关于视觉显著性的计算建模：探究正确之处与遗漏之处。《视觉研究》116（2015），95–112。[8] Alexandre Bruckert、Marc Christie 和 Olivier Le Meur。2023。电影中该看向何处：通过分析视觉注意力理解电影剪辑。《行为研究方法》55（2023），2940–2959。https://doi.org/10.3758/s13428-022- 01949-7[9] L. Chen、Y. Zhang 和 J. Li。2022。视频质量评估的时空注意力机制。《IEEE图像处理汇刊》31（2022），1234 1245 10 1109 2022 3145678 –。https://doi.org/。[10] /TIP.。C. Darcet、J. Lee 和 H. Wang[2024]。。视觉Transformer也需要寄存器。《IEEE模式分析与机器智能汇刊》（2024）。https: //doi.org/ 10 1109 2024 3145679 。/TPAMI.。A. Dosovitskiy、L. Beyer、A. [11] Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani 和 M. Minderer。2020。图像胜过千言万语：面向大规模图像识别的 Transformer。载于国际学习表征会议（ICLR）论文集。OpenReview。https://doi.org/10.48550/arXiv. 2010.11929[12] Richard Droste、Jianbo Jiao 和 J Alison Noble。2020。统一的图像与视频显著性建模。载于欧洲计算机视觉会议。施普林格，419–435。[13] Huiyu Duan、Qiang Hu、Jiarui Wang、Liu Yang、Zitong Xu、Lu Liu、Xiongkuo Min、Chunlei Cai、Tianxiao Ye、Xiaoyun Zhang 等。2025。FineVQ：细粒度用户生成内容视频质量评估。载于计算机视觉与模式识别会议论文集。3206–3217。[14] Myl'ene CQ Farias 和 Welington YL Akamine。2012。增强视觉注意力计算模型的图像质量指标性能研究。《电子快报》48，11（2012），631–633。[15] Pedro Garcia Freitas、Welington YL Akamine 和 Mylène CQ Farias。2018。利用多种时空特征估计视频质量。《信号处理：图像通信》64（2018），1–10。[16] Rohit Girmaji、Siddharth Jain、Bhav Beri、Sarthak Bansal 和 Vineet Gandhi。2025。通过高效解码器和时空动作线索实现简约的视频显著性预测。载于 ICASSP 2025—2025 IEEE国际声学、语音与信号处理会议（ICASSP）论文集。IEEE，1–5。[17] Vlad Hosu、Ting-Chun Lin、Tamas Sziranyi 和 Dietmar Saupe。2020。康斯坦茨自然视频数据库（KonVid-1k）。《IEEE图像处理汇刊》 29 2020 3002 3014 10 1109 2019 2958355 ( ), –。https://doi.org/。[18] /TIP.。Y. Hu[2024] 和 Y. Zhang。。在线处理的视频质量评估：从空间采样到时间采样。《IEEE图像处理汇刊》31（2024），5678 5690 10 1109 2024 3145679 –。https://doi.org/。/TIP.。[19] Yanan Huo[2024] 和 Zhi Li。。基于 Transformer 的多尺度无参考视频质量评估方法。《多媒体系统》30（2024），45–58。https://doi.org/10.1007/s00530-024-01403-y [20] Lai Jiang、Mai Xu、Tie Liu、Minglang Qiao 和 Zulin Wang。2018。DeepVS：基于深度学习的视频显著性预测方法。载于欧洲计算机视觉会议（ECCV）论文集。1–16。https://doi.org/10.1007/ 978-3-030-01258-8 21[21] Tilke Judd、Fredo Durand 和 Antonio Torralba。2009。用于预测人类注视点的显著性计算模型基准测试。载于 MIT技术报告。[22] Jari Korhonen。2019。无参考消费级视频质量评估的两级方法。《IEEE图像处理汇刊》28，12（2019），5923–5938。[23] Solomon Kullback 和 Richard A Leibler。1951。关于信息与充分性。《数理统计年刊》22，1（1951），79–86。https://doi.org/10.1214/ aoms/1177729694[24] Bing Li、Wei Zhang、Mingyi Tian 和 Zhibo Chen。2022。通过质量感知预训练和运动感知盲评野外视频质量。《IEEE视频技术电路与系统汇刊》32（2022），5944–5958。https://doi.org/ 10 1109 2022 3167662 。/TCSVT.。Dingquan Li、Tingting Jiang[2025] 和 Minghua Jiang。2019。野外视频质量评估。载于第 27 届ACM国际会议论文集 2351 2359 [26] Kyle Min 和 Jason J Corso。2019。TASED-Net：用于视频显著性检测的时序聚合空间编码器-解码器网络。载于 IEEE国际计算机视觉大会（ICCV）论文集。2394–2403。[27] Parag K Mital、Tim J Smith、Robin L Hill 和 John M Henderson。2011。动态场景观看中的注视聚类可由运动预测。《认知计算》3，1（2011），5–24。[28] Mayesha Maliha R. Mithila 和 Mylène C. Q. Farias。2025。MS-SCANet：一种用于无参考图像质量评估的多尺度Transformer双重注意力架构。载于 ICASSP 2025 - 2025 IEEE国际声学、语音与信号处理会议（ICASSP）。1–5。https://doi.org/10.1109/ ICASSP49660.2025.10887759[29] A. Mittal、A. K. Moorthy 和 A. C. Bovik。2011。盲/无参考图像空间质量评估器。载于2011年阿西洛马会议：信号、系统与计算机第四十五届会议记录（ASILOMAR）。IEEE，723–727。[30] F. Morra 等。2021。用户生成内容视频质量评估中的变异性。《IEEE图像处理汇刊》30（2021），1123–1134。https: //doi.org/ 10 1109 2021 1234567 . /TIP. . Yunpeng Qu、Kun Yuan、Qizhi [31] Xie、Ming Sun、Chao Zhou 和 Jian Wang。2025。KVQ：通过显著性引导的局部感知提升视频质量评估。载于计算机视觉与模式识别会议论文集。2150–2160。[32] Zhi Ruan 和 Li Jin。2021。利用局部与全局图像特征的时间统计特征进行无参考视频质量评估。《传感器》21，24（2021），8201 10 3390 21248201 。https://doi.org/ . /s Zahra Sinno 和 [33] Alan Conrad Bovik[2018]。大型感知视频质量研究。《IEEE图像处理汇刊》28，2（2018），612–627。https: //doi.org/ 10 1109 2018 2872064 . /TIP. . [34] Charles Spearman[1904]。两种事物之间关联的证明与测量。《美国心理学杂志》15，1（1904），72–101。https: //doi.org/ 10 2307 1412159 . / Wei Sun、Xiongkuo Min、Wei Lu 和 Guangtao Zhai[35]。2022。基于深度学习的用户生成内容视频无参考质量评估模型。载于 ACM国际多媒体会议（ACM MM）论文集。856–865。[36] W. Sun、W. Wen、X. Min、L. Lan、G. Zhai 和 K. Ma。2024。通过设计极简视频质量模型分析视频质量数据集。《IEEE模式分析与机器智能汇刊》（2024）。https://doi.org/10.1109/ TPAMI.2024.1234567[37] Zhi Tu、Yu ming Wang、Neil Birkbeck、Balu Adsumilli 和 Alan C Bovik。2020。UGC-VQA：面向用户生成内容的盲视频质量评估基准测试。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。 10410 10419 10 1109 42600 2020 01043 –。https://doi.org/ . /CVPR . . Z. Tu、Y. Wang、N. Birkbeck、B. Adsumilli[2021] 和 A. C. [38] Bovik。UGC-VQA：面向用户生成内容的盲视频质量评估基准测试。《IEEE图像处理汇刊》30（2021），4449–4464。[39] Laurens van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。《机器学习研究杂志》9，11月（2008），2579–2605。[40] Phuoc V. Vu、Charilaos Christopoulos 和 Damon M. Chandler。2011。用于视频质量评估的时空最显著失真模型。载于 IEEE国际图像处理会议（ICIP）。2505–2508。https: //doi.org/ 10 1109 2011 6116060 . /ICIP. . Wenguan Wang、Jianbing Shen、 [41] Fang Guo、Ming-Ming Cheng 和 Ali Borji。2018。重新审视视频显著性：大规模基准与新模型。载于 IEEE计算机视觉与模式识别会议论文集。4894–4903。[42] Xinyi Wang、Angeliki Katsenou 和 David Bull。2024。ReLaX-VQA：用于提升视频质量评估的残差片段与层堆栈提取。arXiv预印本 arXiv: *2407 11496* 2024 2407 11496 。（）。 [43] https://arxiv.org/abs/ . Zhou Wang[2007] 和 Qiang Li。基于人类视觉速度感知统计模型的视频质量评估。《美国光学学会杂志A辑》24，12（2007），B61–B69。[44] Haoning Wu、Chaofeng Chen、Jingwen Hou、Liang Liao、Annan Wang、Wenxiu Sun、Qiong Yan 和 Weisi Lin。2022。FAST-VQA：通过片段采样实现高效端到端视频质量评估。载于欧洲计算机视觉会议（ECCV）论文集。施普林格，538–554。[45] Haoning Wu、Erli Zhang、Liang Liao、Chaofeng Chen、Jingwen Hou、Annan Wang、Wenxiu Sun、Qiong Yan 和 Weisi Lin。2023。从美学和技术视角探索用户生成内容的视频质量评估。载于 IEEE/CVF国际计算机视觉大会（ICCV）论文集。20144–20154。[46] Haoning Wu、Erli Zhang、Liang Liao、Chaofeng Chen、Jingwen Hou、Annan Wang、Wenxiu Sun、Qiong Yan 和 Weisi Lin。2023。迈向可解释的真实场景视频质量评估：数据库与语言提示方法。载于第31届ACM国际多媒体会议（ACM MM）论文集。 2112 2120 10 1145 3581783 3611737 –。https://doi.org/ . / . [47] Jinjian Wu、Weisi Lin、Sen Deng、Tao Guan、Yuanjie Lin 和 Yuming Wang。2021。用于无参考视频质量评估的长短期卷积Transformer。载于第29届ACM国际多媒体会议论文集

（ 2021 3773 3781 10 1145 3474085 3475368），–。

[48] https://doi.org/ . / . Fengchuang Xing、Yuan-Gen Wang、Hanpin Wang、Leida Li 和 Guopu Zhu。2022。STARVQA：用于视频质量评估的时空注意力。载于*IEEE*国际图像处理会议（*ICIP*）论文集。2326–2330。https: //doi.org/ 10 1109 46576 2022 9897881 . /ICIP . .

[49] Junwen Xiong、Peng Zhang、Tao You、Chuanyue Li、Wei Huang 和 Yufei Zha。2024。Diffsal：用于扩散显著性预测的音视频联合学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。27273–27283。[50] Sheng Yang、Qiuping Jiang、Weisi Lin 和 Yongtao Wang。2019。SGDNet：用于无参考图像质量评估的端到端显著性引导深度神经网络。载于第*27*届*ACM*国际多媒体会议论文集。1383–1391。[51] Zhi Ying、Manav Mandal、Deepayan Ghadiyaram 和 Alan C. Bovik。2021。Patch-VQ：“修补”视频质量问题。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）。14019–14029。 https://doi.org/10.1109/ CVPR46437.2021.01379[52] K. Yuan、H. Liu、M. Li、M. Sun、M. Sun、J. Gong 和 Y. Tang。2024。PTM-VQA：利用来自自然场景的多样化预训练模型进行高效视频质量评估。载于*IEEE/CVF*计算机视觉与模式识别会议论文集

模式识别（*CVPR*）。IEEE，2835–2845。[53] Y. Yuan、Y. Zhang 和 H. Wang。2021。视频质量评估的深度学习：综合综述。*IEEE*视频技术电路与系统汇刊31，5（2021），1623–1636。https://doi.org/10.1109/TCSVT.2020. 2990075[54] Ao-Xiang Zhang、Yuan-Gen Wang、Weixuan Tang、Leida Li 和 Sam Kwong。2023。一种通过全面的人类视觉系统仿真进行时空视频质量评估的方法。*IEEE*控制论汇刊54，8（2023），4749–4762。[55] W. Zhang 和 H. Liu。2017。客观视频质量评估中的显著性研究。《*IEEE*图像处理汇刊》26，3（2017），1275–1288。https://doi.org/ 10 1109 2017 2651410 。/TIP。。X. Zhang、Y. Liu 和 S. Wang。。[56] 2020 使用卷积神经网络进行无参考视频质量评估。《信号处理：图像通信》 86 2020 115 123 10 1016 2020 115123 ( ), –。

[57] 2022 https://doi.org/ 。/j.image. 。Y. Zhao、X. Li 和 Y. Chen。。感知视频质量评估：综合综述。*IEEE*视频技术电路与系统汇刊 32, 4 2022 1234 1245 10 1109 2022 3145678 ( ), –。

[58] 2011 https://doi.org/ 。/TCSVT. 。D. Ćulibrk、M. Mirković、V. Zlokolica、M. Pokrić、V. Crnojević 和 D. Kukolj。。用于视频质量评估的显著运动特征。《*IEEE*图像处理汇刊》20，4（2011），948–958。https://doi.org/10.1109/TIP.2010.2080279