arXiv:2611.32187v1 [cs.MM] 24 Nov 2026

A 用户-项目感知短视频编码框架

Weihui Deng、*He Liu*、*Bin Wang*、*Xiaocheng Li*、*Deliang Fu*、*Zhen Wang*

字节跳动，中国北京

摘要

专业生成内容（PGC）时代的传统编码方法，因用户生成短视频的两大独特特征而面临重大挑战：每日上传量巨大（达到数亿），以及异质化的内容-消费者关系（视频质量各异，消费者所处情境多样）。为应对这些挑战，我们提出了UIAE（用户-项目感知编码），这是一种新型多码率梯度组编码方法，包含三个关键组成部分：1）通过基于规则或基于深度神经网络的模型建立用户-项目关系；2）基于情境消费模式，开发以最大化子群体体验质量（QoE）为目标的局部优化模型；3）通过分层模型集成推导全局最优编码策略。实验结果表明，UIAE框架能够针对具有不同上下文档案的不同用户组，自适应地调整编码结果，从而确保终端观看者获得最优体验质量。我们已将UIAE方法部署到在线流水线中，每日处理数千万个视频转码任务。

索引词— 用户-项目感知编码、短视频、体验质量、最优编码。

1. 引言

传统的端到端视听流水线由若干离散阶段组成：内容上传、转码、内容分发与网络调度（泛 CDN 资源选择[1]和多路径传输[2]），以及播放（客户端预加载和下载调度[3]）（如图1所示）。尤其是转码阶段，需要通过策略性的视频处理生成码率梯度，即一组结构化的视频表示，作为客户端个性化流媒体播放的输入。

面对每天数以亿计的用户生成内容（UGC）处理需求（例如 TikTok/抖音），短视频平台面临着严峻的计算和传输资源限制。转码面临多方面挑战：**i)** 如何在梯级数量有限的情况下最大限度地提升用户体验？实际上，出于投资回报率（ROI）的考虑，大多数视频通常只提供十几个梯级。**ii)** 如何整合用户对视频质量的多样化需求以及

如何将播放流畅度纳入码率梯度决策？例如，如图1所示，iPhone-7 设备在

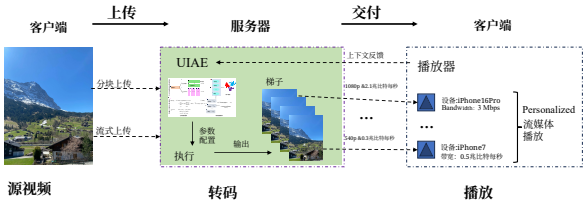

图 1。端到端视听流水线。

网络状况较差时需要低比特率的视频流，而 iPhone-16 设备在网络速度较快时则需要高质量的视频流，以满足对视觉质量的期望。

视频编码优化问题的核心在于策略性地选择压缩参数（例如编解码器标准、编码预设、分辨率和恒定码率因子），以实现最佳的率失真权衡。业界和学术界都提出了多种解决方案，包括：1）固定配置编码[4, 5], 无论内容如何，都采用统一参数。2）自适应编码技术[6–10] 例如内容感知编码（CAE）、逐标题／逐场景／逐镜头／逐块编码，以及上下文感知编码[11–13]。3）基于学习的方法[14–16], 主要优化视频代理指标（例如 VMAF）。4）基于 QoE 的方法[17–27]。

然而，前面三类方法仅关注内容自适应视频处理，未能考虑真实的用户观看体验。与此同时，剩余一类方法对体验质量原则的利用有限，无法满足多样化的用户需求。目前，该领域的研究仍较为匮乏。

为应对这些挑战，本文提出了一种适用于短视频场景的 UIAE 方法。首先，根据相似的播放需求对所有用户群体进行聚类，然后采用基于规则或深度神经网络的方法建立用户-项目关系。接下来，基于上下文档案构建局部优化模型，以最大化各子群体的体验质量，最终形成全局最优编码策略。实验结果表明，与当前最先进的方法相比，UIAE 方法展现出强劲的竞争力。

2. 提出的方法

在本节中，我们提出一种专为以用户生成内容为核心的短视频场景设计的用户-项目感知编码（UIAE）框架，如图2所示。

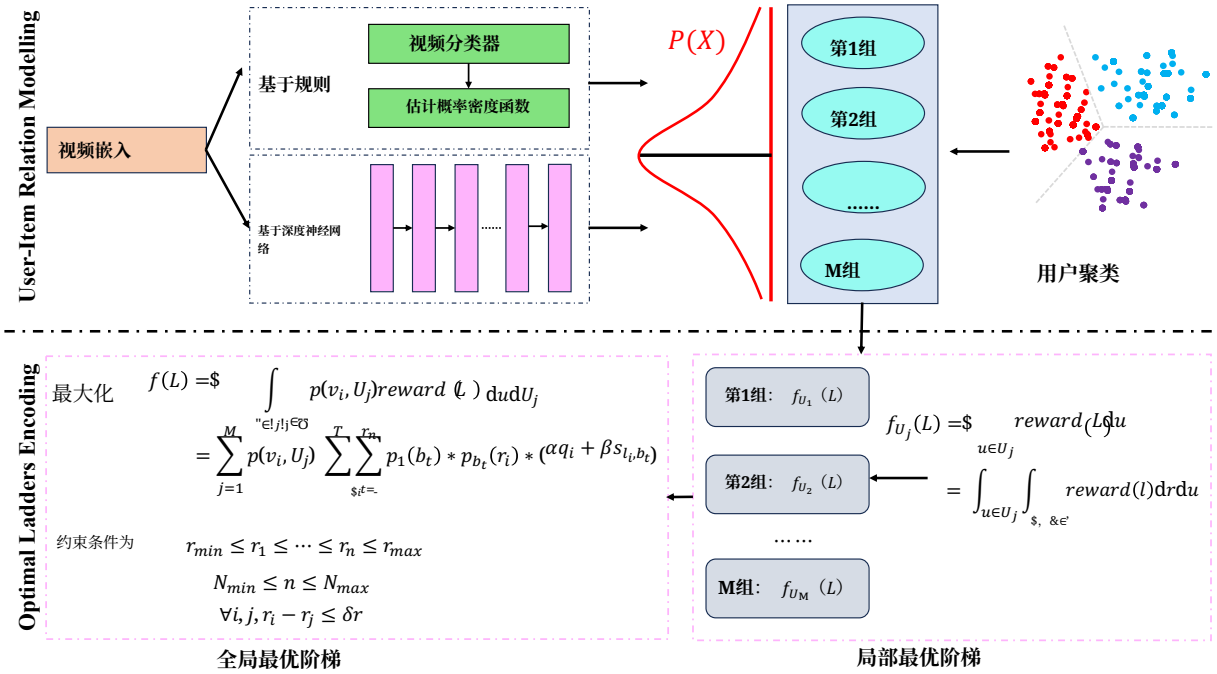

图 2。UIAE 框架。

2.1. 用户-项目关系建模

用户-项目关系模型旨在预测视频上传后的未来播放分布 $P(X)$。我们首先将所有用户划分为若干组，然后根据这些用户组做出转码决策。

用户聚类。用户聚类的目标是将对码率梯度清晰度/流畅度有相似需求的用户划分到同一组。通常，网络状况、设备硬件以及对视频质量和卡顿容忍度的偏好相似的用户，预计也会有相似的码率需求。因此，主要考虑以下特征：1）设备性能指标，包括 CPU/GPU 型号、内存/存储容量等；2）历史播放统计数据，例如网络带宽分布、视频质量/卡顿统计数据，以及不同码率梯度下的观看完成率等。在本研究中，采用 K 均值算法[28]执行聚类任务。最终，所有消费者被划分为 M 个组，即 $\mathcal{U} = \{U_1, U_2, ...U_M\}$。

基于规则的建模。该方法首先使用 CLIP 模型获取视频的多模态嵌入表示[29–31]，然后利用经典的多头分类模型将视频划分为 K 个类别$\mathcal{V} = \{V_1, V_2, ...V_K\}$，其中每个类别代表一种特定的视频类型。随后，计算每个视频类别在 $\{U_1, U_2, ...U_M\}$ 中的播放占比，该占比根据历史播放记录得出。具体而言，用户-项目的播放分布 $P(X)$ 可通过公式1计算。对于 $\forall v_i^* \in V_i$，其在组 v_i^*中的播放概率 U_j 为:

$$p(v_i^*, U_j) = \frac{Play(V_i, U_j)}{Play(V_i, *)} = \frac{\sum_{v_i \in V_i} Play(v_i, U_j)}{\sum_{v_i \in V_i} Play(v_i, *)} \quad (1)$$

其中，$Play(V_i, U_j)$ 表示组 U_j 内所有视频类型为 V_i 的视频的播放次数，$Play(v_i, U_j)$ 表示视频 v_i 在组 U_j 内的播放次数。需要注意的是，$\sum_{j=1}^{M} p(v_i, U_j) = 1$，而 $p(v_i, U_j) \geq 0$。

基于深度神经网络的建模。在基于深度神经网络的建模策略中，模型输入为 $input = \{\mathbf{e_i}, \mathbf{a_i}\}$，其中 $\mathbf{e_i}$ 是视频 v_i 的多模态嵌入，$\mathbf{a_i}$ 表示视频 v_i 的作者侧特征信息。模型输出是一个 M 维向量，表示 M 个组的播放概率，该向量由网络中的 M 个神经元实现。模型的中间结构采用多个简单的全连接层和归一化层，最终的激活函数为 Softmax。我们使用加权交叉熵作为损失函数:

$$\mathcal{L}_{WCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} w_j \cdot y(v_i, U_j) \cdot log(p(v_i, U_j)) \quad (2)$$

其中，w_j是组 U_j的权重，$y(v_i, U_j) \in \{0, 1\}$是一个指示函数，$y(v_i, U_j) = 1$ 表示真实样本中组 U_j 内视频 v_i 的播放概率最高。而 $p(v_i, U_j)$表示深度神经网络模型对组 U_j 内视频 v_i 的播放概率预测值。

2.2. OptimalLadders 编码

最优多码率阶梯组决策过程可分为两部分：1）针对每个

子组进行局部优化；2）针对所有消费者用户进行全局优化。

2.2.1. 局部最优阶梯

给定一个视频 $v*$，我们将多码率阶梯组L 定义为由某个编码决策生成的一组阶梯 $\{l_1, ..., l_n\}$，所有候选阶梯组构成集合 $\mathcal{L} = \{L_1, L_2, ...\}$。每个阶梯 $l_i = (r_i,\ q_i)\ i = 1, ...n$由其比特率 r_i 和视频质量指标 q_i（例如 VMAF）表征。为简便起见，在后续描述中，r_i 可与 l_i 互换使用。阶梯组 L中的比特率按升序排列 $r_1 < r_2 < ... < r_n$。在本研究中，所有候选多码率阶梯均通过调整恒定码率因子、同时固定其他编码器参数生成。于是，在组 U_j 中选择 QoE 最优多码率阶梯组的优化问题可表述如下：

$$f_{U_j}(L) = \int_{u \in U_j} reward(L)\mathrm{d}u = \int_{u \in U_j} \int_{r,l \in L} reward(l)\mathrm{d}r\mathrm{d}u \tag{3}$$

约束条件包括：1）比特率排序约束 $r_{min} \le r_1 \le ... \le r_n \le r_{max}$；2）阶梯数量约束 $N_{min} \le n* \le N_{max}$；3）相邻阶梯之间的比特率差约束 $\forall i, j,\ r_i - r_j \le \delta r$。$reward\ (l)$ 在公式3 中量化了播放阶梯 l时的体验质量。这里，U_j 表示所有可能播放该阶梯组的用户集合 L。通常，我们根据用户的网络带宽和播放概率对用户集合 U_j 进行离散化。具体而言，

$$\begin{aligned}\int_{u \in U_j} reward(L)\mathrm{d}u &= \int_{b_t \in B_j} reward(L)\mathrm{d}b_t \\ &= \sum_{t=1}^{T} \sum_{r_i}^{r_n} p_1(b_t) * p_{b_t}(r_i) * reward(l_i)\end{aligned} \tag{4}$$

其中，$p_1(\cdot)$表示网络带宽概率分布。本文通过分桶对网络带宽进行离散化。$B_j = \{b_1, b_2, ..., b_T\}$ 表示将带宽划分为 T 个桶，$p_{b_t}(r_i)$表示在带宽为 b_t时播放梯级 l_i（比特率为 r_i）的概率。实际上，$p_{b_t}(*)$可以通过拟合历史用户播放数据的统计信息得出。

公式4中的 $reward(l_i)$量化了用户播放梯级 l_i后的体验质量。本文重点关注直接影响用户体验的两个因素：视频清晰度（例如 VMAF）和播放流畅度（例如卡顿时长）。值得注意的是，卡顿时长基于视频比特率与网络带宽的二维分布。令s_{l_i,b_t}表示在网络带宽为 b_t时播放 l_i 的卡顿时长，则：

$$reward(l_i) = \alpha_j * q_i + \beta_j * s_{l_i,b_t} \tag{5}$$

α_j 和 β_j 是超参数，分别表示第 U_j 组用户对视频清晰度和播放流畅度的主观偏好。

最后，公式1 可以离散化为

$$f_{U_j}(L) = \sum_{t=1}^{T} \sum_{r_i}^{r_n} p_1(b_t) * p_{b_t}(r_i) * (\alpha_j * q_i + \beta_j * s_{l_i,b_t}) \tag{6}$$

2.2.2. 全局最优阶梯

基于各组 U_j的局部优化结果，给定视频 $v*$ 的全局最优多码率阶梯组决策问题可形式化表示为：

$$\begin{aligned}\text{maximize} \quad f(L) &= \int_{U_j \subset \mathcal{U}} \int_{u \in U_j} p(v_i, U_j) * reward(L)\mathrm{d}u\mathrm{d}U_j \\ &= \sum_{j=1}^{M} p(v_i, U_j) * \sum_{t=1}^{T} \sum_{r_i}^{r_n} p_1(b_t) * p_{b_t}(r_i) \\ &\quad * (\alpha_j * q_i + \beta_j * s_{l_i,b_t}) \\ \text{满足约束条件}\ & r_{min} \le r_1 \le ... \le r_n \le r_{max} \\ & N_{min} \le n \le N_{max} \\ & \forall i, j, r_i - r_j \le \delta r\end{aligned} \tag{7}$$

显然，这是一个带非线性约束的优化问题，可以采用序列二次规划（SQP）等成熟的优化方法。

3. 实验与讨论

3.1. 实验设置

视频数据集。UIAE框架已部署于抖音的在线转码流程中，每天处理数千万项视频转码任务，为数亿用户提供视频播放服务。为使实验更清晰、简洁，我们从线上生产环境中抽样选取部分视频，构建了验证视频数据集。该数据集包含 5 个不同的内容类别（例如电影/动漫/游戏/休闲视频博客/农业），每个类别包含 10,000 个视频。

用户上下文。用户上下文数据集由从数十亿用户的在线播放数据中收集的性能指标组成，用于生成概率分布 $p_1(\cdot)$、$p_{b_t}(\cdot)$和$s_{l_i,\ b_t}$。在用户聚类阶段，通过比较多种聚类结果确定了 $M = 4$（第1组至第4组）。

评估指标。本研究采用以下指标评估 UIAE 决策框架：平均**比特率**、平均**VMAF**、**卡顿时长**、**卡顿频率**和 **QoE** 评分。前四项指标取自客户端播放事件，而 QoE 评分则根据综合视频质量与卡顿惩罚的线性模型计算得出。例如，$QoE = quality(\cdot) - \rho * stalling\ duration(\cdot) - \gamma * stalling\ frequancy(\cdot)$。我们通过实证配置了 $\rho = 1$，而 $\gamma = 18$ 指的是 [23]。

表1。通过客户端仿真实验对不同方法进行比较分析，结果以所有指标的平均值报告。（Bit：比特率/兆比特每秒，SD：卡顿时长/ms，SF：卡顿频率/%，HP：超参数。 $\boldsymbol{\rho_1} = [0.3, 0.3, 0.3, 0.3]$和 $\boldsymbol{\rho_2} = [0.1, 0.4, 0.5, 1.5]$ 分别表示统一和个性化的参数设置。）

N	方法	HP	QoE	VMAF	Bit	SD	SF
3	CxAE^{quality}	$\boldsymbol{\rho_1}$	56.74	93.79	1.25	25.4	0.65
	CxAE^{QoE}		58.25	93.64	1.16	24.2	0.63
	UIAE$^{QoE_{rule}}$		59.47	93.60	1.17	22.9	0.62
	UIAE$^{QoE_{dnn}}$		**60.34**	93.33	1.13	21.7	0.62
	UIAE$^{QoE_{rule}}$	$\boldsymbol{\rho_2}$	**60.13**	93.24	1.18	22.2	0.61
	UIAE$^{QoE_{dnn}}$		**60.86**	93.13	1.12	21.2	0.60
5	CxAE^{quality}	$\boldsymbol{\rho_1}$	57.13	94.56	1.55	25.5	0.66
	CxAE^{QoE}		58.76	94.42	1.45	24.3	0.64
	UIAE$^{QoE_{rule}}$		59.82	94.33	1.18	23.0	0.63
	UIAE$^{QoE_{dnn}}$		**60.71**	94.20	1.30	21.8	0.63
	UIAE$^{QoE_{rule}}$	$\boldsymbol{\rho_2}$	**60.47**	94.30	1.34	22.3	0.62
	UIAE$^{QoE_{dnn}}$		**61.30**	94.20	1.26	21.3	0.62

3.2. 结果与分析

编码算法比较。我们通过客户端模拟，在视频数据集上比较了 CxAE^{quality}、CxAE^{QoE}、UIAE$^{QoE_{rule}}$ 和 UIAE$^{QoE_{dnn}}$ 的结果。在编码阶段，CxAE^{quality} 和 CxAE^{QoE} 均直接利用整体组的全局上下文。CxAE^{quality} 仅使用网络带宽分布，而 CxAE^{QoE} 则同时纳入网络带宽分布和卡顿分布。UIAE 变体（UIAE$^{QoE_{rule}}$ 和 UIAE$^{QoE_{dnn}}$）采用第1至4组的局部上下文信息，而非直接使用整体组的全局上下文信息。实验比较了两种梯级数量：3 和 5。

表1展示了在梯级数量为 3/5、超参数向量为 $\boldsymbol{\rho_1}$/$\boldsymbol{\rho_2}$时，不同方法的性能。三种基于体验质量的方法均优于面向质量的 CxAE^{quality}。例如，当 N=5 且 HP=$\boldsymbol{\rho_1}$时，UIAE$^{QoE_{dnn}}$ 的体验质量达到 60.71，高于 CxAE^{quality}的 57.13。同时，其卡顿时长为 21.8 ms，卡顿频率为 0.63%，也优于 CxAE^{quality}的 25.5 ms 和 0.66%，因此更符合用户观看需求。在基于体验质量的方法中，两个 UIAE 变体表现更为出色。在五级配置下，这些改进依然存在：体验质量分别提升 1.06 和 1.95，卡顿时长分别减少 1.3 ms（-5%）和 2.5 ms （ -10% ）。总之，两个 UIAE

QoE_{dnn}和 UIAE$^{QoE_{rule}}$ 方法在客户端仿真中展现出更优的服务质量和体验质量表现，尤其在流畅度指标（卡顿时长/频率）方面表现突出，这得益于将用户上下文数据细粒度地整合到多码率阶梯优化目标中。

用户上下文的重要性。 本小节通过分析用户上下文变化对编码结果的影响

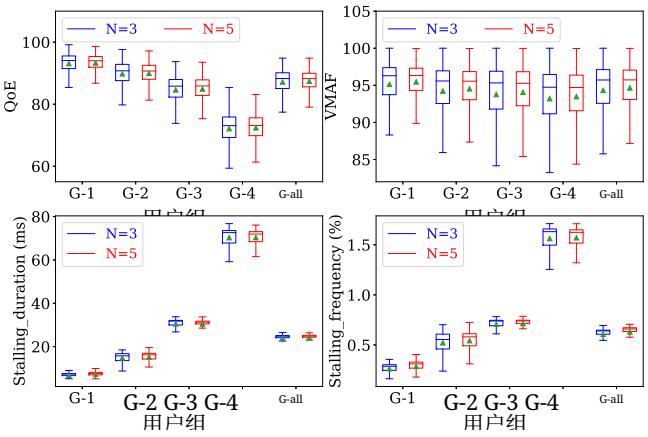

图3。针对具有异构用户上下文的 CxAE^{QoE} 进行比较分析。G-1 至 G-all 分别表示第1组至总体组。超参数为 $\rho = -\frac{\beta_j}{\alpha_j} = 0.3$。

通过使用 CxAE 进行受控实验^{QoE}，采用该方法覆盖所有用户组（网络带宽和卡顿事件从第1组到第4组逐步恶化）。平均仿真结果如 图3所示。第4组的用户上下文档案较差，在所有指标上的表现均有所下降。具体而言，第1至第3组取得了更高的 VMAF分数，卡顿指标显著降低，从而获得了显著更优的体验质量分布。这些结果凸显了用户上下文的关键作用：统一的全局参数无法适应异构的网络带宽和卡顿约束。

分组分析参数。根据第1组至第4组的性能，我们对超参数向量进行了小幅调整 $\boldsymbol{\rho_2}$，具体设置如下：$-\frac{\beta_1}{\alpha_1} = 0.1$，$-\frac{\beta_2}{\alpha_2} = 0.4$，$-\frac{\beta_3}{\alpha_3} = 0.5$，$-\frac{\beta_4}{\alpha_4} = 1.5$。与基线统一设置 $\boldsymbol{\rho_1}$的比较结果如表1所示。UIAE$^{QoE_{dnn}}$ 和 UIAE$^{QoE_{rule}}$均通过面向集群的参数化实现了更高的体验质量（QoE）。具体而言，当梯级数量为 = 3时，UIAE$^{QoE_{dnn}}$ 和 UIAE$^{QoE_{rule}}$ 的 QoE 分别提高了 0.52（+0.86%）和 0.66（+1.11%）。这种个性化优化使模型能够更好地使码率决策与子组特定约束相匹配。

4. 结论

我们提出了 UIAE，一种新颖的多码率梯度优化框架。该框架通过基于规则或基于DNN的方法建立用户-项目关系，随后通过层次化模型集成，将多码率梯决策问题形式化为全局体验质量最大化问题。实验结果证明了 UIAE 的优越性。未来的工作将重点围绕利用多模态行为数据进行个性化偏好建模，以及适用于边缘环境的轻量级自适应展开，逐步构建面向异构用户生成内容生态系统的真正上下文感知流媒体系统。

5. 参考文献

[1] Chunyu Qiao、Tong Liu、Yucheng Zhang 等，《Pira：面向短视频流媒体的泛 CDN 视频内资源自适应》，发表于*ACMMM*，2025 年，第 11987–11995 页。[2] Bowen Hu、Tong Li、Chunyu Qiao 等，《Cross：面向高效多路径视频流媒体的双侧调度框架》，发表于 *EMS*，2025 年，第 13–18 页。[3] Tong Liu、Zhiwei Fan 等，《Deload：基于需求驱动的短视频预加载与可扩展观看时长估计》，发表于*ACMMM*，2025 年，第 6801–6809 页。

[4] 苹果，《苹果设备的 HLS 编写规范》，2023年5月。

[5] 谷歌，《YouTube 推荐的上传编码设置》，2022年。

[6] A. Aaron、Z. Li、M. Manohara、D. Jan 和 D. Ronca，《按标题编码优化》，2015年12月。

[7] Z. Li、A. Aaron、I. Katsavounidis、A. Moorthy 和 M. Manohara，《迈向实用的感知视频质量指标》，2016年6月。

[8]D. Jan、Z. Li、M. Manohara 和 A. Aaron，《云端基于复杂度的一致质量编码》，载于*ICIP*。IEEE，2016年。

[9] I. Katsavounidis，《动态优化器——感知视频编码优化框架》，2018年3月。

[10] C. Chen、Y. Lin、S. Benting 和 A. Kokaram，《利用播放统计数据实现大规模自适应流媒体的优化转码》，发表于*ICIP*。IEEE，2018年。

[11] Y. Reznik、K. Lillevold、A. Jagannath、J. Greer 和 J. Corley，《自适应码率流媒体编码配置的最优设计》，发表于第*23*届 *Packet Video Workshop*，2018年。

[12]Y. Reznik、X. Li、K. Lillevold、A. Jagannath 和 J. Greer，《最优多编解码器自适应码率流媒体》，发表于*ICMEW*。IEEE，2019年。

[13] Y. Reznik、K. Lillevold、A. Jagannath 和 N. Barman，《通过适应内容的“真实分辨率”改进上下文感知编码》，发表于*SMPTE*，2023年。

[14] S. Paul、A. Norkin 和 A Bovik，《通过循环学习进行自适应视频流媒体的凸包预测》，*IEEETIP*，2024年。

[15] V. Menon、P. Rajendran、C.Feldmann、K.Schoeffmann、M. Ghanbari 和 C. Timmerer，《面向自适应直播的感知阈值（JND）感知双遍逐标题编码方案》，*IEEETCSVT*，2023年。[16] M. Francisco、M. Wilmer、G. Juan 和 G. Miguel，《利用深度神经网络进行逐标题和按片段 CRF 估计，适用于基于质量的视频编码》，*ESWA*，2023年。

[17] M. Seufert、S. Egger 等，《HTTP自适应流媒体体验质量综述》，*IEEE* 通信调查与教程，2015年。

[18] P. Lebreton 和 K. Yamagishi，《面向自适应码率视频流媒体的网络与内容相关码率梯度估计》，载于*ICASSP*。IEEE，2021年。

[19] P. Lebreton 和 K. Yamagishi，《面向自适应码率视频流媒体的基于退出率的码率梯度选择机制》，*IEEETMM*，2023年。[20] D. Lorenzi、F. Tashtarian、H. Hellwagner 和 C. Tim- merer，《Medusa：一种用于 HTTP 自适应流媒体的动态编解码器切换方法》，*ACM TOMM*，2024年。

[21] C.Bampis 和 A.Bovik，《学习预测流媒体视频体验质量：失真、重缓冲和记忆效应》，*arXiv*预印本 *arXiv:1703.00633*，2017年。

[22] C. Bampis、Z. Li、I. Katsavounidis 和 A. Bovik，“用于预测流媒体视频体验质量的循环与动态模型”，*IEEE TIP*，2018年。

[23] C. Bampis、Z.Li、I. Katsavounidis 等，“面向感知优化的自适应视频流传输——一个真实的体验质量数据库”，*IEEE TIP*，2021年。

[24] T. Huang、C. Zhang、R.and Wu 和 L. Sun，“利用人类反馈优化自适应视频流传输”，发表于*ACM MM*，2023年。

[25] A. Bentaleb、A. Begen 和 R. Zimmermann，“SDNDASH：利用软件定义网络提升 HTTP 自适应流媒体的体验质量”，发表于*ACM MM*，2016年。

[26] Y. Li、Z. Zhang、H. Chen 和 Z. Ma，“Mamba：将多维自适应码率引入 WebRTC”，发表于*ACMMM*，2023年。

[27] W. Deng、Z. Fan、D. Fu 等，《个性化播放技术：短视频服务如何打造卓越的用户体验》，*arXiv:2410.17073*，2024年。

[28] J. Hartigan 和 M. Wong，《算法 136：一种 k-均值聚类算法》，英国皇家统计学会杂志：*C*辑（应用统计），1979年。

[29] J. Lee、J. Kim、H. Shon、B. Kim、S. Kim、H. Lee 和 J. Kim，《Uniclip：用于对比式语言-图像预训练的统一框架》，*NeurIPS*，2022年。[30]J. Zhou、L. Dong、Z. Gan、L. Wang 和 F. Wei，《非对比式学习与语言-图像预训练相结合》，发表于 *CVPR*，2023年。

[31] B. Wang、K. Zhao、H. Zhao、S. Pu、B. Xiao 和 J. Guo，《视频属性原型网络：零样本视频分类的新视角》，发表于*ICCV*，2023年。