arXiv:2302.13269v1 [cs.CV] 26 Feb 2023

# 基于语义亲和度准则探索无主观意见的视频质量评估

*Haoning Wu*[1] *Liang Liao*[1] *JingwenHou*[1] *Chaofeng Chen*[1] *Erli Zhang*[1]*Annan Wang*[1] *Wenxiu Sun*[2] *Qiong Yan*[2] *Weisi Lin*[1]

**摘要**

近年来，基于学习的视频质量评估（VQA）算法因收集人类质量意见数据的成本高昂而难以实现；此外，由于这些意见存在偏差，算法在各种场景下的鲁棒性也较差。这促使我们探索无主观意见（又称零样本）VQA方法。现有方法仅考虑空间或时间域中的低层自然度，而未考虑高层语义的影响。在本研究中，我们利用对比语言-图像预训练（CLIP）模型中的文本提示，为无主观意见VQA引入显式的语义亲和度指数。我们还通过高斯归一化和 sigmoid 重缩放策略，将其与不同的传统低层自然度指数进行聚合。所提出的盲式统一无主观意见视频质量指数（通过语义与技术指标聚合）（**BUONA-VISTA**）由聚合后的语义和技术指标构成，其性能优于现有无意见感知视频质量评估方法至少**20%**，且比有主观意见的方法更加鲁棒。

## 1. 引言

随着在线视频数量的快速增长，客观视频质量评估日益受到研究人员的关注。近年来，尽管有主观意见的视频质量评估方法 [1–4]已得到广泛探索，但这些方法依赖大量训练数据以及昂贵的人工主观评分 [5–8]，且通常难以适用于新的数据集。如何减轻对高成本训练数据的依赖，并构建能够评估任意给定视频的稳健视频质量评估方法，是亟待研究的问题。

近年来，关于**无意见视频质量评估** [9–12], 的研究寥寥无几，这类方法通常依赖视频质量评估的经验标准，而非基于主观意见数据进行回归。例如，NIQE [9] 通过将图像与 pristine自然内容的分布进行比较，来衡量图像的空间自然度（图1（a））。TPQI [12], 受人类视觉系统相关知识的启发，通过感知域中的帧间曲率来衡量视频的时间自然度 [13, 14]。这些无意见视频质量评估方法

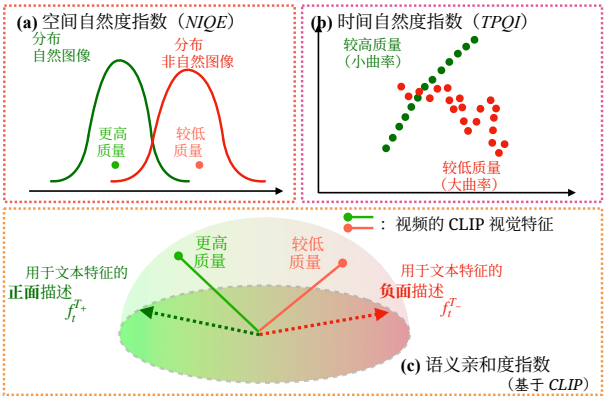



**图 1**：BUONA-VISTA 中三项独立指标准则的可视化

**BUONA-VISTA**。流程如图2所示。

基于低层次标准，忽略了人类对视频语义的感知。此外，许多现有研究注意到，自然真实失真 [15–17] 或与美学相关的问题 [3, 18] 在野外视频中普遍存在，并会影响人类质量感知。这些问题很难通过这些低层次标准捕捉，但利用语义感知深度神经特征可以更好地提取 [4, 19–23]。

本文提出一种语义感知准则，以无监督的方式解决这些高层次质量问题。借助对比语言-图像预训练（CLIP） [24], ，我们能够计算视觉特征与任意给定文本之间的亲和度。基于 CLIP，我们衡量视频的视觉特征与正向（例如，高质量）或负向（例如，低质量）文本描述的文本特征之间的相似程度（图1（c）），以此作为一种语义感知质量准则，主要关注美学相关质量问题以及人类高层次质量感知。借助这些新准则，我们设计了语义亲和度指数，作为一种语义感知的零样本视频质量评估指数，用于评估真实失真和美学问题。此外，我们设计了高斯归一化，随后进行 Sigmoid 重缩放 [25], ，从而将语义亲和度指数与低层次的空间和时间自然度指标加以汇总，并通过语义与技术指标聚合，构成整体的盲式统一无主观意见视频质量指数（**BUONA-VISTA**）。

总的来说，我们的贡献主要有三点：

1) 我们为无意见视频质量评估提出了一种新颖的文本提示式语义亲和度指数。该指数结合了缩写差异亲和度和多提示聚合，

[1]南洋理工大学；　　[2]商汤研究院。

*代码可在此处获取*https://github.com/QualityAssessment/BVQI*。

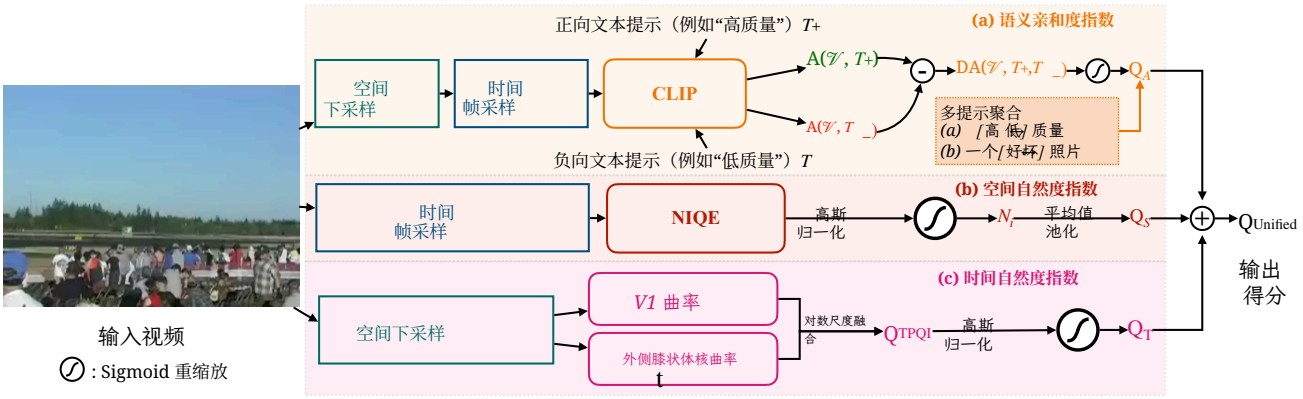



**图 2**：BUONA-VISTA 的整体流程，包括 (a) 语义亲和度指数、(b) 空间自然度指数和 (c) 时间自然度指数。三个指数经过重新映射并聚合，得到最终的 BUONA-VISTA 指数。

以准确匹配人类质量感知。2) 我们提出高斯归一化和 Sigmoid 重缩放策略，将语义亲和度指数与低级空间和时间技术指标对齐并聚合，从而形成 **BUONA-VISTA** 质量指数。3) BUONA-VISTA 的表现显著优于现有零样本视频质量评估指数（>20%），且经证明比有主观意见的视频质量评估方法具有更强的鲁棒性。

## 2. 所提出的方法

在本节中，我们介绍构成所提出的视频质量指数的三项指标，它们采用不同的准则，包括基于 CLIP 的语义亲和度指数（$\mathrm{Q}_A$，第 2.1 节），以及两项技术自然度指标：空间自然度指数（$\mathrm{Q}_S$，第 2.2 节）和时间自然度指数（$\mathrm{Q}_T$，第 2.3 节）。这三项指数经过 Sigmoid 重缩放后，聚合为所提出的**BUONA-VISTA** 质量指数。该指数的整体流程如图 2 所示。

### 2.1. 语义亲和度指数（$\mathrm{Q}_A$）

为提取视频质量评估中的语义相关质量问题，我们使用 CLIP 提取**语义亲和度指数**（$\mathrm{Q}_A$），如下所示。**美学专用数据准备**。由于 BUONA-VISTA 的语义分支旨在处理真实失真和美学问题，而这些问题通常对分辨率或帧率不敏感，因此我们遵循 DOVER 中的数据准备方法 [18]，对原始视频进行空间下采样和时间稀疏帧采样。我们将视频经过下采样后的美学专用视图记为 $\mathcal{V} = \{V_i|_{i=0}^{N}\}$，其中$V_i$ 是 $i$第帧（共采样 $N$ 帧）的下采样视频帧，其空间分辨率为 $224\times224$，与 CLIP 预训练期间采用的空间尺度一致 [24]。

**视频与文本之间的亲和度**。给定任意文本提示$T$，CLIP 中的视觉编码器（$\mathrm{E}_v$）和文本编码器（$\mathrm{E}_t$）会提取$\mathcal{V}$ 和 $T$，将其转换为隐式视觉（$f_{v,\,i}$）和文本（$f_t$）特征：

$$f_{v,i} = \mathrm{E}_v(V_i)|_{i=0}^{N-1}; \qquad f_t^T = \mathrm{E}_t(T) \tag{1}$$

然后，$\mathcal{V}$（视频的美学视角）与文本 $T$ 之间的语义亲和度 $\mathrm{A}(\mathcal{V},\ T)$ 定义如下：

$$\mathrm{A}(\mathcal{V}, T) = \Big(\sum_{i=0}^{N-1} \frac{f_{v,i} \cdot f_t^T}{\|f_{v,i}\|\|f_t^T\|}\Big)/N \tag{2}$$

其中，$\cdot$表示两个向量的点积。

**缩写差异亲和度**。一般而言，质量较好的视频应与**正向**质量相关描述或感受（$T_+$，例如“高质量”“好照片”“清晰”）具有更高的亲和度，而与**负向**质量相关文本描述（$T_-$，例如“低质量”“糟糕的照片”“不清晰”，缩写为 $T_+$）的亲和度较低。因此，我们引入缩写差异亲和度指数（DA），即视频与正向或负向文本的亲和度孰高（图1（c）），将其作为零样本视频质量评估的语义标准：

$$\mathrm{DA}(\mathcal{V}, T_+, T_-) = \mathrm{A}(\mathcal{V}, T_+) - \mathrm{A}(\mathcal{V}, T_-) \tag{3}$$

**多提示聚合**。由于我们希望在语义质量指数中同时提取真实失真（NIQE 或其他低层级指标几乎无法检测到）和审美相关问题，因此我们聚合了两组不同的缩写词对：**1**）高质量↔低质量（$T_{+,\,0}$，$T_{-,\,0}$）；**2**）一张好照片↔一张差照片（$T_{+,\,1}$，$T_{-,\,1}$）。遵循 VQEG 的建议 [25]，，我们进行 Sigmoid 重映射，将两个得分映射到范围 $[0, 1]$（这在实际效果上与人类感知尺度相似），并将重映射后的得分相加，得到最终的语义亲和度指数（$\mathrm{Q}_A$），其形式化定义如下：

$$\mathrm{Q}_A = \sum_{i=0}^{1} \frac{1}{1 + e^{-\mathrm{DA}(\mathcal{V}, T_{+,i}, T_{-,i})}} \tag{4}$$

### 2.2. 空间自然度指数（$\mathrm{Q}_S$）

尽管语义亲和度指数功能强大，我们也使用NIQE [9]指数——首个完全盲测的质量指数——来检测传统类型的**技术失真**，例如加性白高斯噪声（*AWGN*）、*JPEG*压缩

**表 1**：基准评估　　所提出的 BUONA-VISTA 与其他无主观评价质量指标的比较　　.

| 数据集 | LIVE-VQC | | KoNViD-1k | | YouTube-UGC | | CVD2014 | |
|---|---|---|---|---|---|---|---|---|
| 方法 | 斯皮尔曼秩相关系数 | 皮尔逊线性相关系数 | 斯皮尔曼秩相关系数 | 皮尔逊线性相关系数 | 斯皮尔曼秩相关系数 | 皮尔逊线性相关系数 | 斯皮尔曼秩相关系数 | 皮尔逊线性相关系数 |
| **意见感知方法：** | | | | | | | | |
| TLVQM（IEEE图像处理汇刊，2019年）[1] | 0.799 | 0.803 | 0.773 | 0.768 | 0.669 | 0.659 | 0.830 | 0.850 |
| VSFA（ACM多媒体会议，2019年）[3] | 0.773 | 0.795 | 0.773 | 0.775 | 0.724 | 0.743 | 0.870 | 0.868 |
| VIDEVAL（IEEE图像处理汇刊，2021）[20] | 0.752 | 0.751 | 0.783 | 0.780 | 0.779 | 0.773 | 0.832 | 0.854 |
| **现有的无主观意见（零样本）方法：** | | | | | | | | |
| （空间）NIQE（信号处理，2013年）[9] | 0.596 | 0.628 | 0.541 | **0.553** | 0.278 | 0.290 | **0.492** | **0.612** |
| （空间）IL-NIQE（IEEE图像处理汇刊，2015年）[10] | 0.504 | 0.544 | 0.526 | 0.540 | **0.292** | **0.330** | 0.468 | 0.571 |
| （时间）VIIDEO（IEEE图像处理汇刊，2016年）[26] | 0.033 | 0.215 | 0.299 | 0.300 | 0.058 | 0.154 | 0.149 | 0.119 |
| （时间）TPQI（ACM多媒体会议，2022年）[12] | **0.636** | **0.645** | **0.556** | 0.549 | 0.111 | 0.218 | 0.408 | 0.469 |
| **BUONA-VISTA（我们的方法，零样本）** | **0.784** | **0.794** | **0.760** | **0.760** | **0.525** | **0.556** | **0.740** | **0.763** |
| 相较于现有最佳方法的提升 | 23% | 23% | 37% | 38% | 80% | 69% | 50% | 25% |

伪影。这些失真很可能在视频压缩或传输过程中出现。为使不同指数保持一致，我们将归一化原始NIQE分数（$\mathrm{Q}_{\mathrm{NIQE}}$针对 $V_i$）转换为高斯分布 $N$（0，1），并通过Sigmoid重新缩放，以得到逐帧自然度指数（$\mathrm{N}_i$）：

$$\mathrm{N}_i = \frac{1}{1+e^{\frac{\mathrm{Q}_{\mathrm{NIQE},i}-\overline{\mathrm{Q}_{\mathrm{NIQE},i}}}{\sigma(\mathrm{Q}_{\mathrm{NIQE},i})}}} \tag{5}$$

其中，$\overline{\mathrm{Q}_{\mathrm{NIQE}}}$ 和 $\sigma$（$\mathrm{Q}_{\mathrm{NIQE}}$）分别是整个数据集原始 *NIQE* 分数的均值和标准差。接下来，按照 [1, 19, 20], 我们每秒采样一帧（每秒*1*帧），并使用采样帧计算整体**空间自然度指数**（$\mathrm{Q}_S$），$V_{F_k}$ 在 $k$第 - 秒内，计算方式如下：

$$\mathrm{Q}_S = \sum_{k=0}^{S_0} \mathrm{N}_{F_k}/S_0 \tag{6}$$

其中 $S_0$ 是视频的总时长。

### 2.3. 时间自然度指数（$\mathrm{Q}_T$）

虽然 $\mathrm{Q}_A$ 和 $\mathrm{Q}_S$ 能够更好地涵盖不同类型的空间质量问题，但它们无法涵盖时间维度上的失真，例如抖动、卡顿或摄像机运动不平滑；众所周知，这些失真会影响人类质量感知 [1, 27–29] 。一般来说，所有这些时间失真都可以概括为相邻帧之间不平滑的帧间变化，并可通过近期提出的 TPQI [12], 进行捕捉。TPQI 基于连续三帧之间的神经域曲率。具体而言，曲率可通过初级视觉皮层（V1, [13]）和外侧膝状体（LGN， [14]）域上的模拟神经响应计算得出，如下所示：

$$\mathrm{Q}_{\mathrm{TPQI}} = \frac{\log\left(\frac{1}{M-2}\sum_{j=1}^{M-2}\right)\mathbb{C}_j^{\mathrm{V1}} + \log\left(\frac{1}{M-2}\sum_{j=1}^{M-2}\right)\mathbb{C}_j^{\mathrm{LGN}}}{2} \tag{7}$$

由于原始 NIQE/TPQI 得分越低表示质量越好，因此此处（式 $\frac{1}{1+e^{x}}$ 使用 $\frac{1}{1+e^{-x}}$ 而非负向类 sigmoid 重映射（式 5）和式 8）。

其中，$M$ 是整个视频的总帧数，$\mathbb{C}_{\mathrm{LGN}}$ 和 $\mathbb{C}_{\mathrm{V1}}$ 分别是帧 $j$ 处的曲率。随后，通过高斯归一化和 Sigmoid 重缩放，将**时间自然度指数**（$\mathrm{Q}_T$）从原始得分映射得到：

$$\mathrm{Q}_T = \frac{1}{1+e^{\frac{\mathrm{Q}_{\mathrm{TPQI}}-\overline{\mathrm{Q}_{\mathrm{TPQI}}}}{\sigma(\mathrm{Q}_{\mathrm{TPQI}})}}} \tag{8}$$

### 2.4. *BUONA-VISTA* 指数：指标聚合

由于我们的目标是设计一种稳健的无主观意见感知质量指数，因此我们直接将所有指数的尺度对齐得分相加进行聚合，不借助任何视频质量评估数据集进行回归。由于 $\mathrm{Q}_A$、$\mathrm{Q}_S$ 和$\mathrm{Q}_T$ 已分别在式 4、式 5 和式 8 中进行高斯归一化和 sigmoid 重缩放，因此这三个指标的取值范围均为 $[0, 1]$, , 统一后的 **BUONA-VISTA** 指数 $\mathrm{Q}_{\mathrm{Unified}}$ 定义如下：

$$\mathrm{Q}_{\mathrm{Unified}} = \mathrm{Q}_A + \mathrm{Q}_S + \mathrm{Q}_T \tag{9}$$

在下一节中，我们将开展若干实验研究，以证明各个独立指数的有效性以及所提出聚合策略的合理性。

## 3. 实验评估

### 3.1. 实现细节

由于三个指数所针对的质量相关问题各不相同，因此三个分支的输入也有所不同。对于 $\mathrm{Q}_A$，视频会通过双三次 $224 \times 224$下采样核在空间维度上进行下采样，并在时间维度上采样为 [31] 均匀采样帧；对于 Q [18]，视频保留原始空间分辨率，但在时间维度上仅保留均匀采样帧，其中$S_0$ 为视频时长（单位：秒）。对于 Q，所有视频均在空间维度上下采样至 $270 \times 480$ （以保持宽高比），并将所有帧输入神经响应模拟器。该 $\mathrm{Q}_A$ 使用 Python 3.8、PyTorch 1.7 和官方 CLIP-ResNet-50 [32]权重计算。 $\mathrm{Q}_S$ 和 $\mathrm{Q}_T$ 则使用 MATLAB R2022b 计算。

**表2**：比较现有有主观意见方法与零样本BUONA-VISTA（完全无需训练）的跨数据集性能。BUONA-VISTA 显著比这些方法更具鲁棒性。

| 在**训练**于（BUONA-VISTA 为无） | KoNViD-1k | | | | LIVE-VQC | | | | YouTube-UGC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **在测试** | LIVE-VQC | | YouTube-UGC | | KoNViD-1k | | YouTube-UGC | | LIVE-VQC | | KoNViD-1k | |
| | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相 |
| CNN-TLVQM (2020,多媒体) [4] | 0.713 | 0.752 | NA | NA | 0.642 | 0.631 | NA | NA | NA | NA | NA | NA |
| GST-VQA (2021, IEEE 多媒体汇刊) [30] | 0.700 | 0.733 | NA | NA | **0.709** | 0.707 | NA | NA | NA | NA | NA | NA |
| VIDEVAL (2021, IEEE图像处理汇刊) [20] | 0.627 | 0.654 | 0.370 | 0.390 | 0.625 | 0.621 | 0.302 | 0.318 | 0.542 | 0.553 | 0.610 | 0.620 |
| MDTVSFA（2021，IJCV） [23] | **0.716** | **0.759** | **0.408** | **0.443** | 0.706 | **0.711** | **0.355** | **0.388** | **0.582** | **0.603** | 0.649 | 0.646 |
| **BUONA-VISTA (零样本)** | **0.784** | **0.794** | **0.525** | **0.556** | **0.760** | **0.760** | **0.525** | **0.556** | **0.784** | **0.794** | **0.760** | **0.760** |

**表 3**：消融研究（I　　）：所提出的 BUONA-VISTA 中不同指数在三个自然视频数据集上的效果　　.

| BUONA-VISTA 中的不同指数 | | | LIVE-VQC | | KoNViD-1k | | CVD2014 | |
|---|---|---|---|---|---|---|---|---|
| 语义亲和度 ($Q_A$) | 空间自然度 ($Q_S$) | 时间自然度 ($Q_T$) | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ |
| ✓ | | | 0.629 | 0.638 | 0.608 | 0.602 | 0.685 | 0.692 |
| | ✓ | | 0.593 | 0.615 | 0.537 | 0.528 | 0.489 | 0.558 |
| | | ✓ | 0.690 | 0.682 | 0.577 | 0.569 | 0.482 | 0.498 |
| ✓ | ✓ | | 0.692 | 0.712 | 0.718 | 0.713 | 0.716 | 0.731 |
| | ✓ | ✓ | 0.749 | 0.753 | 0.670 | 0.672 | 0.618 | 0.653 |
| ✓ | | ✓ | 0.767 | 0.768 | 0.704 | 0.699 | 0.708 | 0.725 |
| ✓ | ✓ | ✓ | **0.784** | **0.794** | **0.760** | **0.760** | **0.740** | **0.763** |

### 3.2. 评估设置

**评估指标**。 按照常见研究的做法，我们采用两种指标：斯皮尔曼秩相关系数（SRCC），用于评估质量分数与人类意见之间的单调性；以及皮尔逊线性相关系数（PLCC），用于评估线性准确度。

**基准数据集**。 为了更好地评估所提出的 BUONA-VISTA 在不同真实场景设置下的性能，我们选择了四个不同的数据集，包括 CVD2014 [6] (234 个视频，拍摄期间采集了真实失真实验室数据), LIVE-VQC [8](585 个由智能手机录制的视频), KoNViD-1k [7] (1200 个从社交媒体平台收集的视频),以及 YouTube-UGC [17，33](1131 个可用视频，其中包含从 YouTube 收集的非自然视频，类别包括 游戏/动画/歌词视频）。

### 3.3. 基准比较

我们在表 1/2 中，将现有的无主观意见（零样本）或有主观意见视频质量评估方法进行比较，以评估所提出的 BUONA-VISTA 指数的准确性和稳健性。

**与无意见感知方法的比较**。 所提出的 BUONA-VISTA 质量指数显著优于所有现有的无主观意见质量指数，提升至少 ***20%*** 。具体而言，在三个自然场景的视频质量评估数据集（LIVE-VQC、KoNViD-1k 和 CVD2014）上，其皮尔逊线性相关系数/斯皮尔曼秩相关系数几乎达到 0.8，甚至可与某些有主观意见方法相媲美，或表现得更好。在非自然场景数据集（YouTube-UGC）上，借助语义亲和度指数，所提出的 BUONA-VISTA 相较于所有语义无关零样本质量指数取得了惊人的**80%** 提升，首次 在该数据集上提供了合理的质量预测。无需任何训练，




**图 3**：从三个独立指数及总体 BUONA-VISTA 指数的角度展示质量最佳/最差的视频。所有演示视频均附于补充材料中。

这些结果表明，所提出的 BUONA-VISTA 相较于现有指标实现了跨越式提升，可广泛用作稳健的真实场景视频质量指标。

**与意见感知方法的比较**。 尽管 BUONA-VISTA 几乎不可能超越有主观意见的方法，但它已在很大程度上弥合了零样本方法与监督方法之间的差距。此外，这些有主观意见的方法还可能面临在特定数据集上过拟合的额外挑战。将其跨数据集性能与表2 中所提出的 零样本 BUONA-VISTA 的结果进行比较可以发现，当训练集和测试集并非来自同一组视频和意见时，BU ONA-VISTA 能够超越现有的基于观点的方法，进一步证明了所提出 BUONA-VISTA 的鲁棒性。

**表4**：消融研究 （IV）：不同文本提示及所提出的多提示聚合策略的影响 .

| | BUONA-VISTA 的整体性能 | | | 仅使用语义亲和度指数时的性能 | | | |
|---|---|---|---|---|---|---|---|
| 数据集 | LIVE-VQC | KoNViD-1k | CVD2014 | LIVE-VQC | KoNViD-1k | CVD2014 | YouTube-UGC |
| **提示词对** | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ |
| **(a)** *[*高 ↔低*]* 质量 | 0.768/0.775 | 0.725/0.725 | 0.738/0.757 | 0.560/0.575 | 0.477/0.472 | **0.728/0.729** | 0.539/0.564 |
| **(b)** *a [*好↔坏*]* 照片 | 0.778/0.785 | 0.727/0.727 | 0.653/0.686 | 0.608/0.581 | 0.586/0.551 | 0.507/0.512 | 0.473/0.458 |
| **(a)+(b)** 聚合 | **0.784/0.794** | **0.760/0.760** | **0.740/0.763** | **0.629/0.638** | **0.609/0.602** | 0.686/0.693 | **0.585/0.606** |

**表 5**：消融研究（II）：所提出的 BUONA-VISTA 在 YouTube-UGC 数据集上不同指数的影响。

| BUONA-VISTA 中的指数 | | | YouTube-UGC | |
|---|---|---|---|---|
| $Q_A$ | $Q_S$ | $Q_T$ | 斯皮尔曼秩相关系数↑ | 皮尔逊线性相关系数↑ |
| ✓ | | | 0.585 | **0.606** |
| ✓ | ✓ | | **0.589** | 0.604 |
| ✓ | ✓ | ✓ | 0.525 | 0.556 |
| | ✓ | | 0.240 | 0.153 |
| | | ✓ | 0.133 | 0.141 |

**表6**：消融研究（III）：所提出的 BUONA-VISTA 中不同聚合策略的比较。

| **聚合** | LIVE-VQC | KoNViD-1k | CVD2014 |
|---|---|---|---|
| 指标 | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ | 斯皮尔曼秩相关系数↑/皮尔逊线性相关系数↑ |
| 直接相加 | 0.760/0.750 | 0.675/0.660 | 0.664/0.699 |
| 线性 +加法 | 0.776/0.760 | 0.720/0.710 | 0.700/0.729 |
| *Sigmoid* +乘法 | 0.773/0.729 | 0.710/0.679 | 0.692/0.661 |
| *Sigmoid* +加法 | **0.784/0.794** | **0.760/0.760** | **0.740/0.763** |

### 3.4. 定性研究

在定性研究中，我们分别可视化每个指数以及 BUONA-VISTA 总体指数中得分最高或最低的视频快照。如图3所示，（**a**）语义亲和性与美学高度相关，其中（**b**）空间自然度关注空间纹理（清晰↔模糊），而（**c**）时间自然度关注时间变化（稳定↔抖动），这与上述三个指数的评价标准相一致。我们还在补充材料中附上了这些示例的原始视频。

### 3.5. 消融研究

在消融研究中，我们讨论不同质量指数的影响：语义亲和度、空间自然度和时间自然度（第3.5.1节），以及它们对自然度的影响。随后，我们讨论不同聚合策略的影响（第3.5.2节）。此外，我们还评估不同提示词对以及所提出的多提示聚合方法的影响（第3.5.3节）。

#### *3.5.1.* 不同指数的影响

**自然数据集上的评估**。在评估各个指数的单独影响时，我们将四个数据集分为两部分：第一部分将 LIVE- VQC、KoNViD-1k 和 CVD2014 归类为**自然数据集**，因为它们不包含计算机生成内容，也不包含类似电影的剪辑和拼接视频。不同设置的结果列于表3中，三个指数均对最终结果有显著贡献

从而证明，与语义相关的质量问题、传统的空间失真和时间失真对于构建能够稳健估计人类质量感知的模型都很重要。具体而言，在视频仅包含拍摄过程中产生的真实失真的 CVD2014 中，语义亲和度（$Q_A$）指数的贡献最大；在普遍认为存在最多时间失真的数据集 LIVE-VQC 中，时间自然度（$Q_T$）指数对整体准确率的贡献最大。这些结果证实了我们前述的观点，即这三个指数分别关注不同方面。

**在 YouTube-UGC 上的评估**。在 YouTube-UGC 上，如表5所示，空间自然度指数无法提升 BUONA-VISTA 的最终性能，而时间自然度指数甚至导致性能下降 8%。由于 YouTube-UGC 中的视频时长均较长（20 秒），且几乎每个视频都由多个场景组成，我们推测这种性能下降可能源于场景转换期间的情况：此时，时间曲率很大，但并不意味着质量下降。在未来的工作中，我们将考虑检测视频中的场景转换，并仅在同一场景内计算时间自然度指数。

#### *3.5.2.* 聚合策略的影响

我们通过与不同的重缩放策略进行比较，评估表 6 中聚合策略的影响（线性 表示仅进行高斯归一化，而 *Sigmoid* 表示先进行高斯归一化，再进行 Sigmoid 重缩放），以及不同的融合策略（加法（+）或乘法（×））。结果表明，高斯归一化和 Sigmoid 重缩放都有助于提升聚合指标的最终性能，并且加法 优于乘法。

#### *3.5.3.* 不同文本缩写词对的影响

在表4中，我们讨论了不同文本缩写对作为 $T_+$ 和 $T_-$ 在公式3中的影响。我们注意到，*[*高↔低*]* 质量在 CVD2014 上表现非常出色，此时可以忽略内容多样性，主要关注的是**真实失真**。然而，对于 LIVE-VQC 和 KoNViD-1k（具有多样化的美学特征），*[*好↔差*]*照片提示的准确率更高。结果表明，不同数据集关注的质量问题各不相同，而聚合两组缩写对能够稳定提升所有数据集的整体性能，证明了所提出的多提示聚合策略的有效性。

## 4. 结论

本文提出了 BUONA-VISTA，这是一种面向野外视频的鲁棒零样本无主观意见视频质量指数，它在空间和时间维度上对基于 CLIP 文本提示的语义亲和力指数与传统技术指标进行了对齐和聚合。所提出的 BUONA-VISTA 在无主观意见视频质量指数中实现了前所未有的性能，并且相较于有意见的视频质量评估方法，在不同数据集上展现出更强的鲁棒性。我们希望，所提出的鲁棒视频质量指数能够成为相关视频研究中可靠且有效的指标，并为实际应用作出贡献。

## 5. 参考文献


[1] J. Korhonen，“用于无参考消费级视频质量评估的双层方法”，IEEE 图像处理汇刊，第28卷，第12期，第5923–5938页，2019年。1，3[2] H. Wu、C. Chen、J. Hou、L. Liao、A. Wang、W. Sun、Q. Yan 和 W. Lin，“Fast-VQA：采用片段采样的高效端到端视频质量评估”，载于欧洲计算机视觉会议，2022年。1[3] D. Li、T. Jiang 和 M. Jiang，“野外视频的质量评估”，载于ACM 多媒体会议，2019年，第2351–2359页。1，3[4] J. Korhonen、Y. Su 和 J. You，“通过统计时间特征和深度空间特征进行盲自然视频质量预测”，载于ACM 多媒体会议，2020，第3311–3319页。1，4[5] Z. Ying、M. Mandal、D. Ghadiyaram 和 A. Bovik，“Patch-VQ：‘修补’视频质量问题”，载于IEEE/CVF 计算机视觉与模式识别会议，2021年。1[6] M. Nuutinen、T. Virtanen、M. Vaahteranoksa、T. Vuori、P. Oit- tinen 和 J. H¨akkinen，“CVD2014——用于评估无参考视频质量评估算法的数据库”，IEEE 图像处理汇刊，第25卷，第7期，2016年。1，4[7] V. Hosu、F. Hahn、M. Jenadeleh、H. Lin、H. Men、T. Szir´anyi、S. Li 和 D. Saupe，“康斯坦茨自然视频数据库（KoNViD-1k）”，载于多媒体体验质量评估会议，2017年，第1–6页。1，4[8] Z. Sinno 和 A. C. Bovik，“感知视频质量的大规模研究”，IEEE 图像处理汇刊，第28卷，第2期，第612–627页，2019年。1，4[9] A. Mittal、R. Soundararajan 和 A. C. Bovik，“构建‘完全盲’图像质量分析器”，IEEE 信号处理快报，第20卷，第3期，第209–212页，2013年。1，2，3[10] L. Zhang、L. Zhang 和 A. C. Bovik，“一种特征丰富的完全盲图像质量评估器”，IEEE 图像处理汇刊，第24卷，第8期，2015年。1，3[11] A. Mittal、A. K. Moorthy 和 A. C. Bovik，“空间域中的无参考图像质量评估”，IEEE 图像处理汇刊，第21卷，第12期，2012年。1[12] L. Liao、K. Xu、H. Wu、C. Chen、W. Sun、Q. Yan 和 W. Lin，“探索视频感知表征在盲视频质量评估中的有效性”，载于ACM 多媒体会议，2022年。1，3[13] F. Tong，“初级视觉皮层与视觉意识”，《自然综述：神经科学》，第4卷，第3期，第219–229页，2003年。1，3[14] D. H. O’Connor、M. M. Fukui、M. A. Pinsk 和 S. Kastner，“注意调节人类外侧膝状体中的反应”，《自然-神经科学》，第5卷，1209，2002 11，1203 1，3 no. pp. – 。

[15] Y. Fang、H. Zhu、Y. Zeng、K. Ma 和 Z. Wang，“智能手机摄影的感知质量评估”，载于IEEE/CVF 计算机视觉与模式识别会议。[16] Z. Ying、H. Niu、P. Gupta、D. Mahajan、D. Ghadiyaram 和 A. Bovik，“从图像块到图片（paq-2-piq）：映射图像质量的感知空间”，载于IEEE/CVF 计算机视觉与模式识别会议，2020 1 。J. G. Yim、Y. [17] Wang、N. Birkbeck 和 B. Adsumilli，“YouTube 用户生成内容数据集的主观质量评估”，载于国际图像处理会议，2020。1，4[18] H. Wu、L. Liao、C. Chen、J. Hou、A. Wang、W. Sun、Q. Yan 和 W. Lin，“解耦用户生成内容视频质量评估中的美学效应和技术效应”，2022年。1，2，3[19] H. Wu、C. Chen、L. Liao、J. Hou、W. Sun、Q. Yan、J. Gu 和 W. Lin，“用于高效端到端视频质量评估的邻域代表性采样”，2022 1, 3 。Z. Tu、Y. Wang、[20] N. Birkbeck、B. Adsumilli 和 A. C. Bovik，“UGC-VQA：面向用户生成内容的盲视频质量评估基准”，IEEE 图像处理汇刊，第30， 4449 4464, 2021 1, 3, 4 pp. – 。W. Zhang、K. Ma、[21] J. Yan、D. Deng 和 Z. Wang，“使用深度双线性卷积神经网络进行盲图像质量评估”，IEEE视频技术电路与系统汇刊，第30， 1， 36 47, 2020 1 no. pp. – 。B. Li、[22] W. Zhang、M. Tian、G. Zhai 和 X. Wang，“通过质量感知预训练和运动感知盲评估野外视频质量”，IEEE视频技术电路与系统汇刊，2022 1 。D. Li、T. Jiang 和 [23] M. Jiang，“采用混合数据集训练的野外视频统一质量评估”，国际计算机视觉杂志，第129， 4, 2021 1, 4 [24] no. 。A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever，“ 从自然语言监督中学习可迁移视觉模型 ”，2021。1，2 [25] “视频质量专家组关于视频质量评估客观模型验证的最终报告”，视频质量专家组。[26] A. Mittal、M. A. Saad 和 A. C. Bovik，“完全盲视频完整性预言机”，IEEE 图像处理汇刊，第25， 1， 289 300, 2016 3 no. pp. – 。J. You 和 J. [27] Korhonen，“用于无参考视频质量评估的深度神经网络”，载于国际图像处理会议，2019 3 。W. [28] Kim、J. Kim、S. Ahn、J. Kim 和 S. Lee，“深度视频质量评估器：从时空视觉敏感性到卷积神经聚合网络”，载于欧洲计算机视觉会议，2018 3 。H. Wu、[29] C. Chen、L. Liao、J. Hou、W. Sun、Q. Yan 和 W. Lin，“DiscovQA：用于视频质量评估的时间失真内容变换器”。[30] B. Chen、L. Zhu、G. Li、F. Lu、H. Fan 和 S. Wang，“学习用于无参考视频质量评估的广义时空深度特征表示”，IEEE视频技术电路与系统汇刊，2021 4 。R. Keys，“用于 [31] 数字图像处理的三次卷积插值”，IEEE 声学、语音与信号处理汇刊，第29， 6， 1153 1160, 1981 3 no. pp. – 。K. [32] He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于IEEE/CVF 计算机视觉与模式识别会议，2016， 770 778 3 pp. – 。Y. [33] Wang、S. Inguva 和 B. Adsumilli，“用于视频压缩研究的YouTube 用户生成内容数据集”，载于2019年多媒体信号处理研讨会，2019年。4