arXiv:2602.03892v1 [cs.CV] 3 Feb 2026

分割后审核：面向语言指代的音视分割的无参考掩码质量评估

Jinxing Zhou[1], Yanghao Zhou[2],Yaoting Wang[3],Zongyan Han[1],Jiaqi Ma[1],Henghui Ding[3], Rao Muhammad Anwer[1],Hisham Cholakkal[1]
[1]MBZUAI, [2]新加坡国立大学, [3]复旦大学

摘要

语言指代的音视分割（*Ref-AVS*）旨在通过联合推理视频、音频和文本，对自然语言描述的目标对象进行分割。除了生成分割掩码之外，如何提供丰富且可解释的掩码质量诊断仍鲜有研究。在本文中，我们引入了*Ref-AVS*场景下的掩码质量评估（*MQA-RefAVS*），这是一项新任务，旨在不依赖推理阶段的真实标注作为参考的情况下，评估候选分割掩码的质量。给定音视频语言输入和每个提供的分割掩码，该任务要求估计其与未观测真实标注之间的交并比，识别相应的错误类型，并推荐可执行的质量控制决策。为支持该任务，我们构建了*MQ-RAVSBench*，这是一个涵盖多样且具有代表性的掩码错误模式的基准测试，涉及几何问题和语义问题。我们进一步提出了*MQ-Auditor*，这是一种基于多模态大语言模型（*MLLM*）的审核者，能够显式推理多模态线索和掩码信息，从而生成定量和定性的掩码质量评估。大量实验表明，*MQ-Auditor*优于强大的开源和商业多模态大语言模型，并且可以与现有*Ref-AVS*系统集成，以检测分割失效并支持后续的分割改进。数据和代码将发布于 *https://github.com/jasongief/MQA-RefAVS*。

1. 引言

物体分割一直是计算机视觉中的基础问题，经历了从图像分割 [43]到视频物体分割 [81]的发展。随着外部文本和音频引导的引入，研究进一步扩展到参考引导的视频物体分割 [10]和音频引导的视觉分割 [29]。在此基础上，语言指代的音视分(Ref-AVS) [54] 近期被提出，并成为多模态分割领域的热门研究主题。给定一段带有同步音频的视频和一段参考文本，Ref-AVS 任务旨在生成分割掩码，以在音视场景中准确定位指代对象。

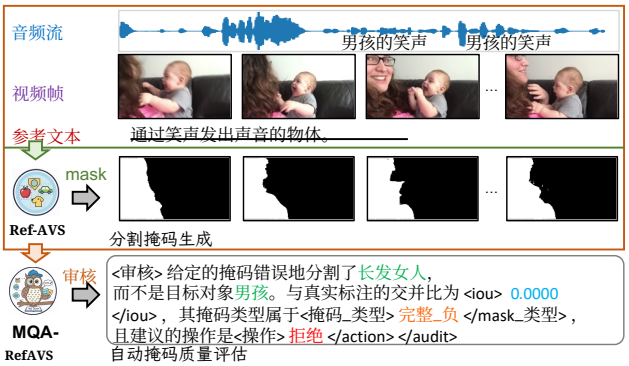


图 1. 任务示意图。以往的 Ref-AVS 方法旨在分割目标对象。相比之下，我们提出的 MQA-RefAVS 任务侧重于自动掩码质量评估，从而识别掩码错误，并为进一步细化提供合适的操作。

当前 Ref-AVS 研究几乎都聚焦于生成分割掩码，并主要通过计算这些掩码与真实标注掩码之间的交并比（IoU）来评估模型性能。然而，在实际部署中，往往无法获得真实标注掩码（即无参考）。例如，在构建具有开创性的 Ref-AVSBench 数据集 [54] 以用于 Ref-AVS 任务，或在新数据集上评估 Ref-AVS 分割模型时，目标数据的真实标注掩码都不可用。在这种情况下，审核者（人或模型）不仅必须解读音频、视频和参考文本以识别目标对象，还必须仔细判断预先标注或生成的掩码是否确实对应预期对象，以及是否足以用于下游

✉通信邮箱：*jinxing.zhou@mbzuai.ac.ae*。

使用。这一过程与分割本身有本质区别，因此构成了一个独立问题：**掩码质量评估（MQA）**。在实际应用中，分割系统不可避免地会生成存在各种失败模式的掩码。图 1 展示了一个示例：此前最先进的 Ref-AVS 模型 TGS-Agent [79]，未能推理多模态内容，错误地分割了女人，而非目标男孩。此前，我们只能依靠真实标注掩码获得单一的交并比得分。与仅总结性能的标量评估指标相比，掩码质量评估能提供更丰富、更易解读的评估结果（见图 1），从而支持细粒度的质量控制。识别出生成的掩码中的质量问题后，掩码质量评估可支持重新生成掩码或进行针对性修订，从而构建更可靠的分割流程。然而，现有数据集和模型几乎无法支持对这一环节进行显式建模或自动化。

为弥补这一空白，我们提出 Ref-AVS 场景下的掩码质量评估（MQA-RefAVS），这是一项无需访问真值标注、自动推断候选分割掩码质量的新任务。如图 1 所示，给定多模态输入和每帧预测的掩码，该任务需要估计掩码交并比、识别其错误类型，并推荐适当的质量控制操作。为系统研究多样且可控的失败模式，我们构建了 MQ-RAVSBench，这是首个专为 Ref-AVS 掩码质量评估设计的基准测试。MQ-RAVSBench 基于 Ref-AVSBench 数据集 [54] 构建，包含 1,840 个视频和 2,046 条参考文本，并由此生成 26,061 个掩码实例。对于每个 <视频、参考文本 > 对，我们生成六种具有代表性的掩码类型，包括完美、剪切、膨胀、腐蚀、合并和全负，涵盖多种失败模式。每个掩码都标注了其交并比，并附有一项推荐操作，可选接受、小幅修订、重大修订或拒绝。这些掩码共同模拟了真实的质量变化，范围从完全准确的预测，到轻微的几何瑕疵，再到严重的语义错误。所有掩码均使用开源多模态模型和工具自动构建，详细流程见第 3 节。

此外，我们提出了**MQ-Auditor**，这是一种基于多模态大语言模型（MLLM）的审核者，能够明确推理音频、视觉、语言和掩码信息。MQ-Auditor通过监督式指令微调进行训练。与以往方法不同，MQ-Auditor评估掩码质量并估算交并比，无需在推理期间使用真值掩码，因此可直接应用于部署阶段。在MQ-RAVSBench上的大量实验表明，对于通用多模态分割任务而言，掩码质量评估仍然颇具挑战，尤其是在多模态Ref-AVS场景中，无论是开源还是商业多模态大语言模型，包括 Gemini-3-Flash [17]，都难以胜任。相比之下，MQ-Auditor始终能提供更准确、可靠的质量评估。此外，MQ-Auditor还能有效补充现有的Ref-AVS模型，识别分割失败并提升分割性能。

总而言之，我们的贡献有三点：**1)**我们将掩码质量评估确定为语言指代的音视分割中一个全新且重要的问题。**2)** 我们建立了 MQ-RAVSBench，这是首个用于系统性掩码质量评估的数据集。**3)** 我们提出了 MQ-Auditor，这是一种多模态审核器，能够在没有真实标注的情况下推断掩码质量，并支持下游分割改进。

2. 任务：MQA-RefAVS

我们在一项具有代表性的多模态分割任务——Ref-AVS 设定中探讨掩码质量评估（MQA）。具体而言，给定一个视频 $\mathcal{V}$ 及与之同步的音 $\mathcal{A}$、一个描述目标对象的指代表达 $\mathcal{R}$、第 t 个片段（$t \in [1, T]$）中的关键视频帧 $\mathcal{V}_t$，以及相应的候选二值掩码 $\mathcal{M}_t$，MQA-RefAVS 任务要求审核模型 Φ通过预测以下内容来评估掩码质量：1）候选掩码与未观测到的真实标注掩码之间的交并比（IoU） $\boldsymbol{s}$，作为定量质量指标，其 $s \in [0, 1]$为；2）掩码类型 $\boldsymbol{m}$，从六种预定义类别中选择：完美、全负、剪切、膨胀、腐蚀、合并；3）用于质量控制的建议操作 $\boldsymbol{a}$，从接受、小幅修订、重大修订、拒绝中选择。第 3.2 节将介绍掩码类型和操作类别的详细信息。为简洁起见，我们省略 s、 m 和 a的下标t 。总之，MQA-RefAVS 任务可表述为：

$$s, m, a = \Phi(\mathcal{V}, \mathcal{A}, \mathcal{R}, \mathcal{V}_t, \mathcal{M}_t). \tag{1}$$

这项任务需要联合感知音频、视频和指代语言，以识别目标对象，判断其是否已由给定帧中的候选掩码准确分割，并最终给出定量和定性评估。

3. 数据集：MQ-RAVSBench

为促进所提出任务的开展，我们提出了 MQ-RAVSBench。下面介绍其构建细节。

3.1. 数据来源与划分

我们从现有的 Ref-AVSBench 数据集中获取视频 [54]，最初为 Ref-AVS 任务提出。Ref-AVSBench 中的视频时长为 10 秒，且每个视频均关联多条指代文本，可能针对不同的对象。对于每个 <视频、参考> 对，该

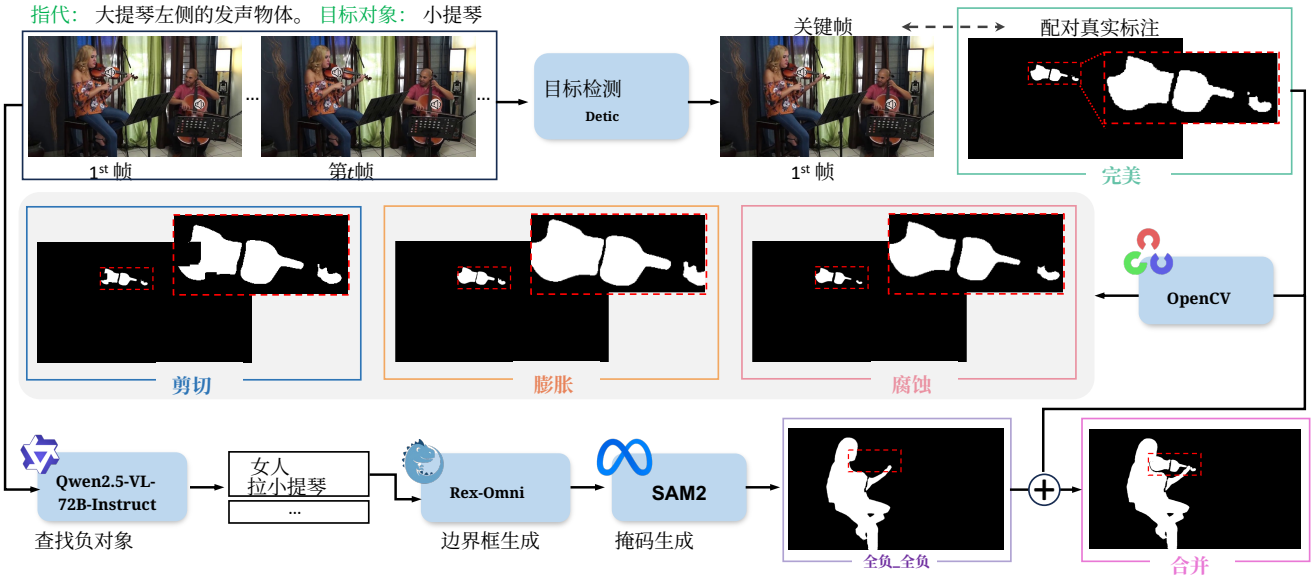


图2。MQ-RAVSBench 的掩码构建流程。对于训练集和基于图像的评估，我们采用目标检测模型 Detic [82], 来识别包含最丰富物体集合的关键帧。基于真实标注**完美**的 Ref-AVSBench 掩码 [54], 我们使用 OpenCV 库生成具有几何质量问题的掩码，包括**剪切**、**膨胀**和**腐蚀**。此外，我们使用强大的多模态大语言模型/视觉语言模型构建流程，生成**全负**掩码，这些掩码对应完全错误的物体，并存在严重的语义质量问题。通过合并全负和完美掩码，我们得到**合并**掩码。

已提供指代表达的物体类别，以及针对10个均匀采样视频帧的二值分割掩码。这些元数据对我们的数据集构建至关重要。在视频采样过程中，我们仔细平衡了视频唯一性、指代文本模板和对象类别。因此，MQ-RAVSBench包含1,840个视频，其中1,306个用于训练，534个用于测试。为便于未来对开放词汇与零样本泛化进行评估，测试集被进一步划分为两个子集：269个视频属于已见集，265个视频属于未见集，划分依据是对象类别是否出现在训练集中。每个训练视频配有一条指代文本，从而得到1,306<视频、参考> 实例。对于测试集，每个视频可关联一条或两条指代文本，已见子集和未见子集分别得到437个和303个实例。值得注意的是，每个训练和测试实例都用质量等级各不相同的多种掩码（7∼13）进行了增强（见第3.2节），从而大幅扩展了整体数据集规模（见表1）。

3.2. 掩码分类体系与质量标注

在确定视频与指代文本对后，我们进一步确定具有不同质量等级的对应分割掩码，计算 IoU 值，并定义推荐操作。具体而言，我们考虑六种掩码类型，以模拟人工标注和真实场景中的模型预测中常见的情况。构建流程概览如图2所示。

完全准确（完美）。候选掩码 $\mathcal{M}_t$ 准确地分割了目标指代对象，精确勾勒出其形状，包含所有真正例像素，并与真实标注掩码 $\mathcal{G}_t$完美匹配。此类掩码直接取自 Ref-AVSBench [54], , 其中每个实例都包含一个完美掩码。因此，此类掩码的 IoU 值 s为 1，建议操作a 为接受：

$$\mathcal{M}_t = \mathcal{G}_t \Rightarrow m = \{\text{perfect}\}, s = 1, a = \{\text{accept}\}. \quad (2)$$

完全错误（全负）。在这种情况下，候选掩码 $\mathcal{M}_t$ 相对于真实标注完全错误，分割出了完全无关的物体或背景区域。当人工标注者或分割模型误解音视频语言线索并选择了错误目标时，常会出现此类错误。 $\mathcal{M}_t$ 中的所有像素均为假正例，因此交并比值为0，建议的操作是拒绝：

$$\mathcal{M}_t \cap \mathcal{G}_t = 0 \Rightarrow m = \{\text{full_neg}\}, s = 0, a = \{\text{reject}\}. \quad (3)$$

我们使用自动化流程生成完整负掩码。对于每个<视频，给定已知的目标对象类别、参考>实例及相应的视频帧 $\mathcal{V}_t$, 我们采用强大的视觉语言模型 Qwen2.5-VL-72B-Instruct-AWQ [2], 生成最多五个负对象候选项。模型不会只输出简单的物体类别，而是提供描述性名词短语，从而更准确地区分视觉上相似的物体。具体提示词见 G.1 节。

我们利用这些负对象短语，借助 Rex-Omni [21], ——一种支持指代表达定位的先进多模态大语言模型——来定位相应的边界框。随后将这些边界框用作 SAM 2 [47]的提示，以生成分割掩码。为进一步筛选具有挑战性的负样本，我们计算剩余负掩码与真实标注掩码之间的边界框交并比，并保留重叠度最高的三个掩码。边界框交并比越高，表明负对象在空间上越接近目标对象或与其发生交互，因此生成的掩码对多模态理解和质量评估更具挑战性。如图 2 所示，示例中的目标对象是小提琴，而生成的负对象是拉小提琴的女性。值得注意的是，在较简单的场景中，有效的完整负掩码可能少于三个，数量从零到三个不等。内部挖除（剪切）。在这种情况下，候选掩码$\mathcal{M}_t$成功定位了目标对象，但遗漏了其内部区域中的部分正像素 $\mathcal{C}_t$（图 2 ）。此类错误可能源于分割模型的能力有限（参见先前研究中的图 8 和图 9 [77]）。此外，分割模型通常采用固定阈值来获取最终掩码，阈值设置的差异也可能导致不同区域的像素被移除。考虑到这些因素，我们通过将生成结果的交并比限制在两个范围 $[0.85, 0.9]$ 和 $[0.75, 0.8]$内，分别生成两个剪切掩码。这些范围是通过初步实验确定的，可生成既真实又对质量评估具有挑战性的掩码。交并比较高的掩码被视为困难样本，其中的挖除错误更细微、更难评估；而交并比较低的范围对应中等难度。因此，困难样本和中等难度样本的建议操作分别定义为小幅修订和重大修订。基于 Ref-AVSBench 提供的真实值 Perfect 掩码，我们使用 OpenCV[1] 库，通过不同尺寸的矩形或椭圆形结构元素生成剪切掩码。该过程可概括如下：

$$\mathcal{M}_t = \mathcal{G}_t \setminus \mathcal{C}_t, \mathcal{C}_t \subset \mathcal{G}_t \Rightarrow m = \{\text{cutout}\}, \quad s \in (0,1), a = \{\text{minor revision; major revision}\}. \quad (4)$$

局部膨胀（膨胀）。虽然候选掩码 $\mathcal{M}_t$成功覆盖了目标对象的所有像素，但其边界错误地扩展到相邻区域 $\mathcal{D}_t$，造成向外过度分割，包含了背景像素。与剪切掩码类型类似，我们分别生成具有高和中难度级别的两种膨胀掩码。相应的建议操作为小幅修订和重大修订，定义如下：

$$\mathcal{M}_t = \mathcal{G}_t \cup \mathcal{D}_t, \mathcal{D}_t \subset \text{neighbor}(\mathcal{G}_t) \Rightarrow m = \{\text{dilate}\}, \quad s \in (0,1), a = \{\text{minor revision; major revision}\}. \quad (5)$$

局部腐蚀（腐蚀）。在这种情况下，候选掩码 $\mathcal{M}_t$ 中的所有像素都属于目标对象，但该掩码未能覆盖对象边界附近的像素 $\mathcal{E}_t$，导致向内欠分割。遵循与剪切和膨胀类型相同的策略，我们构建了两种腐蚀掩码，难度级别分别为高和中，对应的操作为小幅修订和重大修订：

$$\mathcal{M}_t = \mathcal{G}_t \setminus \mathcal{E}_t, \mathcal{E}_t \subset \text{boundary}(\mathcal{G}_t) \Rightarrow m = \{\text{erode}\}, \quad s \in (0,1), a = \{\text{minor revision; major revision}\}. \quad (6)$$

与非目标对象合并（合并）。在此情况下，候选掩码 $\mathcal{M}_t$ 正确分割了目标对象，但也错误地过度分割了其他干扰对象 $\mathcal{H}_t$。在实践中，我们通过将前文介绍的完美掩码与全负掩码合并来构造此类掩码。因此，每个实例最多可生成三个合并掩码。随后，我们计算生成掩码的交并比，并根据交并比值定义建议操作。由于纳入了不同的负对象，合并掩码的交并比 s 可能处于 **(0，1)** 范围内。在我们的设置中，我们进一步将此范围划分为三种情况：**1）**如果$s \in [0.9, 1)$，正确分割的目标对象占合并掩码的比例较大，负对象的影响有限；我们将建议操作定义为小幅修订（困难样本）；**2）**如果 $s \in [0.75, 0.9)$，其中包含了较大比例的负像素，建议操作为重大修订（中等难度样本）；**3）**如果 $s \in$ **(0，0.75)**，建议操作为拒绝（简单样本）。我们采用相对较高的阈值（例如，0.75），因为即使合并掩码的交并比较高，它仍可能错误地包含一个不同的对象，这表明存在严重的语义误解，应予以明确惩罚。此过程可概括如下：

$$\mathcal{M}_t = \mathcal{G}_t \cup \mathcal{H}_t, \mathcal{G}_t \cap \mathcal{H}_t = \varnothing \Rightarrow m = \{\text{merge}\}, \quad s \in \textbf{(0，1)}\ a = \{\text{小幅修订；重大修订；拒绝}\}. \quad (7)$$

总之，我们构造掩码以反映广泛的常见情形，包括准确分割（完美）、欠分割（剪切、腐蚀、全负）和过度分割（膨胀、合并）。这些掩码类型共同涵盖了不同程度的几何与语义质量问题，并呈现出多样的交并比分布及相应的质量控制操作。

3.3. 训练与评估协议

训练协议。训练集包含 1,306 < 个视频、参考>个实例。在我们的设置中，对于每个训练实-

[1]https://github.com/opencv/opencv

表1。MQ-RAVSBench 的统计信息。我们展示了不同训练集和测试集划分中视频、指代表达和掩码样本的分布。基于图像的 和基于视频的评估分别表示测试是在每个视频的单个关键帧或所有帧上进行。

		#视频	#指代表达。	#掩码
训练集		1,306	1,306	16,761
基于图像的评估				
测试集（已见）		269	437	5,609
测试集（未见）		265	303	3,691
	总计	534	740	9,300
基于视频的评估				
测试集（已见）		60	60	7,674
测试集（未见）		40	40	4,966
	总计	100	100	12,640

例，我们按照第 3.2 节选取一个代表性视频帧，以构建六种掩码类型。我们采用强大的目标检测模型 Detic [82],，在以 1 帧/秒采样的 10 个视频帧中检测潜在对象，并选取检测到对象数量最多的帧。这样的关键帧包含丰富的视觉内容，因此在构建 全负和合并掩码时，能够支持更多样的负对象选择。因此，训练集包含 16,761 个掩码样本。这种基于单帧的设计提高了训练效率，同时使模型能够从多样化的对象类别、掩码类型、交并比值和动作标签分布中学习。

评估协议。 如表 1所示，我们在测试时考虑两种评估协议： ***1***）基于图像的评估。 与训练阶段类似，对于每个视频，我们在单个关键帧上评估模型的掩码质量评估（MQA）性能。已见和未见测试集中的掩码数量分别为5,609和3,691。该协议评估模型的基础掩码质量评估能力，包括多模态语义理解、IoU估计准确率以及建议操作的正确性。 ***2***）基于视频的评估。 在此协议中，对每个视频的全部10帧进行评估。为便于分析并降低推理阶段的耗时，我们分别从已见和未见测试集中选取60个和40个视频，并要求同一视频中的所有帧具有相同的掩码类型。此设置产生了7,674个和4,966个测试掩码样本。与基于图像的协议相比，基于视频的评估还能够分析模型对掩码质量预测的时间一致性。各掩码类型的详细统计数据见第 8节 B。

评估指标。 在定义训练和测试方案后，我们确定评估指标。作为引言——
如第 2 节所述，该任务需要预测交并比 s、掩码类型 m和建议操作 a。因此，我们设

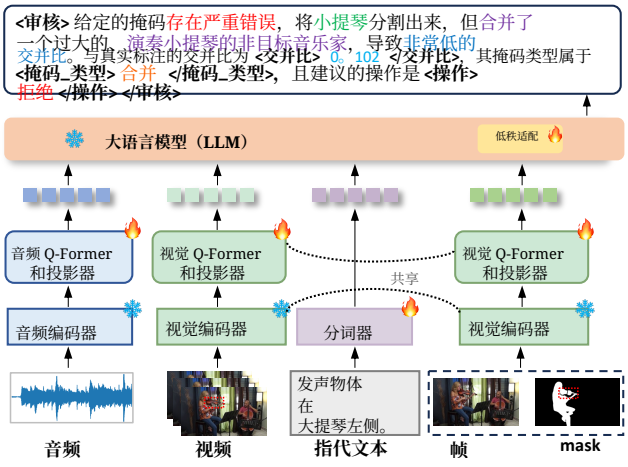


图 3. 我们的 MQ-Auditor 模型示意图。

计了用于评估模型在这些方面表现的指标：
• **均方根误差**。该指标计算所有评估样本中预测交并比与 s 真值交并比之间的均方根误差（RMSE）（见公式8）。

• ***F_2*-分数**。掩码类型和动作预测均使用 F_β 分数进行评估， $\beta = 2$ （见公式9）。这种选择突出了召回率，这对于掩码质量评估系统而言是理想的，原因有二：1）漏掉有问题的掩码，通常比错误地标记正确掩码代价更高；2）我们的实证结果表明，掩码质量评估模型往往能达到较高的精确率（通常接近 100%），但召回率相对较低。各评估协议的详细计算见第C节。

4. 方法：MQ-Auditor

网络。为了解决 MQA-RefAVS 任务，我们提出了一种名为 MQ-Auditor 的基线方法。该方法利用多模态大语言模型（MLLM）分析多模态信号，并以自然语言生成判断结果，从而进行掩码质量评估。如图3所示，音频 $\mathcal{A}$和视频 $\mathcal{V}$ 分别由对应模态的编码器 BEATs [5] 和 CLIP-ViT-L/14 [44], 进行处理。随后，提取的特征被输入带有 32 个可学习查询词元的 Q-Former 模块 [28]，再经线性投影器处理，得到音频和视觉潜在嵌入。指代表达 $\mathcal{R}$ 经过词元化，转换为文本词元嵌入。具体而言，给定一个原始视频帧 $\mathcal{V}_t$ 和一个候选掩码 $\mathcal{M}_t$，我们首先通过对 $\mathcal{V}_t$ 和 $\mathcal{M}_t$进行逐元素相乘，生成一个掩码帧 $\mathcal{V}'_t$。与原始帧 $\mathcal{V}_t$相比，掩码帧 $\mathcal{V}'_t$ 能够明确突出掩码选定的区域。二值掩码 $\mathcal{M}_t$ 还会通过通道复制转换为伪 RGB 图像，并与 $\mathcal{V}'_t$拼接。所得表示由视频编码所用的同一视觉编码器和投影层进行处理。最后，对应的多模态嵌入

表2。在 MQ-RAVSBench 上进行的基于图像的评估。我们将 MQ-Auditor 与强大的开源及商业模型进行比较，包括 Video-LLaMA3-7B [68], Qwen2.5-Omni-7B [60], Ming-Flash-Omni [1], 和 Gemini-3-Flash-Preview [17]。 F_2-M（%）和 F_2-A（%）分别表示掩码类型和动作预测的 F_2-分数。“H”和“M”分别表示困难样本和中等困难样本。“All”表示评估该类型的所有掩码样本。“平均值”表示所有掩码类型的平均性能。最佳和次佳结果分别以 **粗体** 和下划线标出。

设置	指标	方法	完美	剪切		膨胀		腐蚀		合并	全负	Avg.
			All	H	M	H	M	H	M	All	All	
Seen	均方根误差 ↓	Video-LLaMA3-7B	**0.249**	0.268	0.291	0.283	0.313	0.272	0.301	0.573	0.957	0.390
		Qwen2.5-Omni-7B	0.315	0.342	0.359	0.351	0.365	0.354	0.352	0.541	0.904	0.432
		Ming-Flash-Omni	0.997	0.875	0.810	0.859	0.761	0.860	0.762	0.613	0.543	0.787
		Gemini-3-Flash-Preview	0.399	0.363	0.338	0.327	0.297	0.347	0.308	**0.304**	**0.251**	0.326
		MQ-Auditor（我们的方法）	0.359	**0.199**	**0.160**	**0.298**	**0.232**	**0.329**	**0.305**	0.334	0.507	**0.303**
	F_2-M ↑	Video-LLaMA3-7B	86.76	0.57	0.00	0.00	0.00	0.00	0.00	0.20	0.42	9.77
		Qwen2.5-Omni-7B	**88.32**	10.91	14.71	0.00	0.00	0.00	0.00	0.72	12.35	14.11
		Ming-Flash-Omni	2.29	6.24	5.68	0.00	0.00	0.29	0.29	24.13	18.21	6.35
		Gemini-3-Flash-Preview	20.57	77.78	79.00	0.57	3.13	3.41	3.99	73.98	**80.40**	38.09
		MQ-Auditor（我们的方法）	60.58	**83.41**	**91.78**	**59.92**	**78.39**	**36.09**	**47.15**	**74.55**	69.43	**66.81**
	F_2-A ↑	Video-LLaMA3-7B	**89.48**	0.86	0.00	0.57	0.00	0.29	0.00	0.00	0.00	10.13
		Qwen2.5-Omni-7B	88.12	11.46	0.00	7.05	0.00	5.66	0.00	4.96	12.35	14.40
		Ming-Flash-Omni	1.43	0.86	4.84	1.43	3.71	0.57	3.99	64.21	68.91	16.66
		Gemini-3-Flash-Preview	20.57	20.04	62.12	23.97	38.05	**24.74**	**47.11**	**68.32**	**88.39**	43.70
		MQ-Auditor（我们的方法）	60.58	**29.35**	**85.78**	**24.74**	**71.57**	20.04	44.08	43.00	74.52	**50.41**
未见	均方根误差 ↓	Video-LLaMA3-7B	**0.192**	0.259	0.306	0.264	0.298	0.254	0.289	0.524	0.951	0.371
		Qwen2.5-Omni-7B	0.418	0.501	0.455	0.431	0.412	0.394	0.370	0.501	0.916	0.489
		Ming-Flash-Omni	0.981	0.886	0.843	0.850	0.757	0.860	0.761	0.653	0.518	0.790
		Gemini-3-Flash-Preview	0.409	0.357	0.334	0.320	0.298	0.345	0.314	**0.301**	**0.250**	0.325
		MQ-Auditor（我们的方法）	0.274	**0.161**	**0.135**	**0.203**	**0.154**	**0.257**	**0.269**	0.310	0.508	**0.252**
	F_2-M ↑	Video-LLaMA3-7B	**88.28**	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.16	9.92
		Qwen2.5-Omni-7B	74.74	24.71	26.19	0.00	0.00	0.00	0.00	1.13	3.71	14.50
		Ming-Flash-Omni	5.52	15.85	14.60	0.00	0.00	0.43	1.71	26.63	17.64	9.15
		Gemini-3-Flash-Preview	18.39	78.07	79.40	0.86	3.16	3.14	3.97	73.84	**80.42**	37.92
		MQ-Auditor（我们的方法）	64.34	**93.29**	**97.07**	**76.22**	**89.43**	**27.30**	**41.26**	**77.73**	70.92	**70.84**
	F_2-A ↑	Video-LLaMA3-7B	**94.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	10.47
		Qwen2.5-Omni-7B	74.44	25.08	0.00	14.81	0.00	13.64	0.00	6.94	3.71	15.40
		Ming-Flash-Omni	5.94	1.36	11.34	2.99	9.69	2.99	8.86	57.88	66.20	18.58
		Gemini-3-Flash-Preview	21.98	19.78	64.31	23.29	37.47	24.85	**47.15**	**69.08**	**88.49**	44.04
		MQ-Auditor（我们的方法）	64.34	**47.76**	**76.82**	**37.05**	**81.20**	**28.77**	30.60	44.57	74.41	**53.95**

随后 $<\mathcal{A},\ \mathcal{V},\ \mathcal{R},\ \mathcal{V}_t,\ \mathcal{M}_t,\ \mathcal{V}'_t>$ 被注入预定义的系统提示（第G.2节），并输入大语言模型主干网络 LLaMA-2-7B-Chat [53]。如图3所示，大语言模型的输出包括自然语言质量分析、交并比估计，以及对掩码类型和建议操作的预测（更多指令微调提示见第G.3节）。

训练。音频/视觉-文本对齐的预训练阶段按照 [79]进行。随后，我们使用参数高效的 LoRA 层进行指令微调，秩为 32，缩放因子为 64。批次大小设为 4。训练集包含 1,306 个视频，产生16,761 <视频、参考、掩码> 样本（表1），其中完美掩码仅占很小一部分。为平衡正负样本，我们采用一种采样策略：每个训练轮次包含 1,306 个样本，每个视频选取一个掩码，并强制每个小批次中的两个完美掩码（正样本比例 $p = 50\%$）。该策略使模型能够在多个训练轮次中逐步观察到全部 16,761 个样本。在实践中，我们使用 bf16 精度在四块 NVIDIA A100 GPU（40GB）上将 MQ-Auditor 训练 48 个训练轮次。我们使用 AdamW 优化器，初始学习率为 1×10^{-4}。

评估。对于基于图像的评估协议，MQ-Auditor 使用单个关键帧及其对应的候选掩码处理每个视频。均方根误差和 F_2-分数通过对所有帧级样本（公式10&11）的结果取平均值计算。对于基于视频的评估协议，MQ-Auditor 同时考虑所有视频帧的预测结果。首先汇总帧级预测，以获得视频级均方根误差和 F_2-分数，然后对所有测试视频取平均值（公式12&13）。我们将发布训练和评估代码，以促进结果的可复现性。

表3. MQ-Auditor在MQ-RAVSBench上的基于视频的评估。对于合并和完全负类，我们主要评估掩码交并比最高的困难样本。 F_2-M和 F_2-A分别表示掩码类型和动作预测的 F_2分数。

设置	指标	完美	剪切		膨胀		腐蚀		合并	全负	Avg.
		All	H	M	H	M	H	M	H	H	
Seen	均方根误差 ↓	0.276	0.163	0.148	0.231	0.180	0.279	0.251	0.243	0.556	0.258
	F_2-M ↑	57.02	86.17	93.24	61.68	75.64	29.78	42.85	51.16	55.77	61.48
	F_2-A ↑	57.02	34.92	84.57	21.24	70.76	11.76	40.02	40.99	58.34	46.63
未见	均方根误差 ↓	0.185	0.199	0.183	0.190	0.170	0.190	0.222	0.299	0.567	0.245
	F_2-M ↑	67.72	87.26	94.27	57.86	82.62	28.35	39.44	53.83	49.47	62.31
	F_2-A ↑	67.72	45.83	76.91	25.31	70.04	26.30	29.08	45.61	52.37	48.80

表 4. 掩码信息利用方式的消融研究。对于剪切、膨胀和腐蚀掩码类型，我们首先分别评估困难（H）和中等困难（M）样本，并报告其平均结果。“平均值”表示所有列的平均值。

指标	策略	Seen						未见						Avg.
		完美	剪切	膨胀	腐蚀	合并	全负	完美	剪切	膨胀	腐蚀	合并	全负	
均方根误差 ↓	仅掩码 ($\mathcal{M}_t$)	**0.294**	0.186	**0.248**	**0.287**	0.357	0.588	0.304	0.176	0.238	0.306	0.364	0.604	0.329
	掩码帧 ($\mathcal{V}'_t$)	0.523	0.304	0.341	0.453	0.367	**0.383**	0.431	0.249	0.300	0.400	0.351	**0.407**	0.376
	两者 ($[\mathcal{M}_t; \mathcal{V}'_t]$)	0.359	**0.180**	0.265	0.317	**0.334**	0.507	**0.274**	**0.148**	**0.179**	**0.263**	**0.310**	0.508	**0.303**
F_2-M ↑	仅掩码（$\mathcal{M}_t$）	**75.32**	86.25	38.44	32.57	59.63	63.84	**77.41**	**95.32**	63.00	24.46	59.98	63.28	61.62
	掩码帧（$\mathcal{V}'_t$）	56.12	71.88	59.47	20.40	72.55	**81.78**	54.74	78.98	72.72	15.39	76.02	**81.78**	61.82
	两者（$[\mathcal{M}_t;\ \mathcal{V}'_t]$）	60.58	**87.60**	**69.15**	**41.62**	**74.55**	69.43	64.34	95.18	**82.83**	**34.28**	**77.73**	70.92	**69.02**
F_2-A ↑	仅掩码（$\mathcal{M}_t$）	**75.32**	56.46	28.13	26.65	36.36	67.46	**77.41**	**64.10**	38.73	20.33	33.39	66.30	49.22
	掩码帧（$\mathcal{V}'_t$）	56.12	48.47	42.85	18.71	39.11	**85.51**	54.74	51.81	49.20	21.69	39.27	**84.33**	49.32
	两者（$[\mathcal{M}_t;\ \mathcal{V}'_t]$）	60.58	**57.57**	**48.16**	**32.06**	**43.00**	74.52	64.34	62.29	**59.13**	**29.68**	**44.57**	74.41	**54.19**

5. 实验

5.1. 主要结果

我们通过将 MQ-Auditor 与若干最先进的开源和闭源多模态大语言模型进行比较，对 MQ-RAVSBench 进行基准测试，其中包括 Video-LLaMA3-7B [68],Qwen2.5-Omni-7B [60], Ming-Flash-Omni [1], 和 Gemini-3-Flash-Preview [17]。值得注意的是，Ming-Flash-Omni 采用稀疏混合专家（MoE）架构，总参数量为 100B，每个 token 仅激活 6.1B 参数。所有参与比较的模型都能够直接处理音频和视频输入并进行推理，因此比较公平且具有参考价值。除非另有说明，我们均采用基于图像的评估协议进行比较。定量结果汇总于表2。我们发现，Video-LLaMA3 和 Qwen2.5-Omni 倾向于接受大多数候选掩码，因此在完美掩码类型上表现出色，但在负掩码类型上的结果明显较弱。相比之下，Ming-Flash-Omni 表现得过于保守，经常拒绝候选掩码，导致其在完美掩码上的表现不佳，但在全负类型上的结果相对较好。能力更强的 Gemini-3-Flash 在不同掩码类型上表现得相对均衡，尤其擅长识别全负掩码。尽管如此，在已见和未见设置下取得了最佳的总体平均值（“平均值”）。定性比较见第***E*** 节（图 ***5***∼***10***）。MQ-Auditor 的推理效率也很高（见表11）。这些结果表明，尽管现有的多模态大语言模型在通用音频-视频理解方面展现出强大的能力，例如字幕生成和问答，但对于专业且细粒度的 MQA-RefAVS 任务而言，它们仍然不够完善。表 3 进一步报告了 MQ-Auditor 在基于视频的评估协议下的性能，从而完成了我们的基准测试分析。

5.2. 消融研究

本节报告基于图像的评估协议下的消融研究结果。

掩码利用的影响。 我们的方法首先将候选掩码 $\mathcal{M}_t$ 应用于原始视频帧 $\mathcal{V}_t$ ，得到掩码帧 $\mathcal{V}'_t$，然后将其与原始掩码拼接 $\mathcal{M}_t$ ，以提取供 MQ-Auditor 使用的输入嵌入。我们通过考察两种变体来研究掩码用法的影响：仅使用掩码 $\mathcal{M}_t$ 或仅使用掩码帧 $\mathcal{V}'_t$。如表 4所示，仅使用掩码有助于评估完美 和 剪切 类型，因为二值掩码提供了直接的几何线索。相比之下，仅使用掩码帧对全负类型尤其有效，因为掩码帧清晰地呈现了掩码所捕获的语义内容。结合这两种表征，模型便能利用互补的几何与语义信息，从而取得最佳的整体性能。

表5。IoU估计策略的消融研究。“直接NTP”表示通过直接的下一词元预测生成IoU值。“回归ST”表示首先生成一个特殊词元（例如，<IoU值>），随后使用线性层回归IoU。

指标	策略	Seen						未见						Avg.
		完美	剪切	膨胀	腐蚀	合并	全负	完美	剪切	膨胀	腐蚀	合并	全负	
均方根误差 ↓	回归ST	0.758	0.680	0.590	0.612	0.474	**0.226**	0.565	0.618	0.459	0.471	0.412	**0.298**	0.514
	直接NTP	**0.359**	**0.180**	**0.265**	**0.317**	**0.334**	0.507	**0.274**	**0.148**	**0.179**	**0.263**	**0.310**	0.508	**0.303**
F_2-M ↑	回归ST	**69.03**	85.83	58.75	34.52	60.46	68.77	**74.44**	93.95	78.00	**36.04**	58.49	**71.51**	65.82
	直接NTP	60.58	**87.60**	**69.15**	**41.62**	**74.55**	**69.43**	64.34	**95.18**	**82.83**	34.28	**77.73**	70.92	**69.02**
F_2-A ↑	回归ST	**69.03**	54.09	40.32	27.16	40.34	70.66	**74.44**	55.99	50.99	28.78	36.72	72.56	51.76
	直接NTP	60.58	**57.57**	**48.16**	**32.06**	**43.00**	**74.52**	64.34	**62.29**	**59.13**	**29.68**	**44.57**	**74.41**	**54.19**

表 6. 模型训练中每批正样本比例的消融研究 p。

指标	比例 p	Seen						未见						Avg.
		完美	剪切	膨胀	腐蚀	合并	全负	完美	剪切	膨胀	腐蚀	合并	全负	
均方根误差 ↓	25%	0.409	0.200	0.280	0.348	0.353	**0.420**	0.378	0.173	0.270	0.304	0.345	**0.440**	0.327
	50%	0.359	**0.180**	**0.265**	0.317	**0.334**	0.507	0.274	**0.148**	**0.179**	0.263	**0.310**	0.508	**0.303**
	75%	**0.268**	0.207	0.286	**0.310**	0.362	0.652	**0.156**	0.168	0.187	**0.239**	0.368	0.716	0.327
F_2-M ↑	25%	40.24	**88.47**	**77.23**	**51.69**	**77.94**	**76.99**	52.77	94.91	**83.21**	**46.30**	**78.52**	**77.05**	**70.44**
	50%	60.58	87.60	69.15	41.62	74.55	69.43	64.34	**95.18**	82.83	34.28	77.73	70.92	69.02
	75%	**86.37**	81.43	36.89	12.40	60.31	55.17	**89.99**	92.17	53.60	14.01	48.85	49.50	56.72
F_2-A ↑	25%	40.24	**60.32**	**53.17**	**39.71**	**44.51**	**82.26**	52.77	**66.00**	54.78	**37.95**	44.15	**80.55**	**54.70**
	50%	60.58	57.57	48.16	32.06	43.00	74.52	64.34	62.29	**59.13**	29.68	**44.57**	74.41	54.19
	75%	**86.37**	52.18	26.36	10.98	32.18	59.62	**89.99**	59.77	35.62	16.00	27.75	50.92	45.64

表7. 先前 Ref-AVS 分割模型（EEMC [54] 和 TGS-Agent [79]）在应用 MQ-Auditor 进行掩码质量评估前后的表现。

方法	Seen			未见		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
EEMC	31.9	42.2	37.1	16.7	32.5	24.6
+ MQ-Auditor	**38.4**	**50.6**	**44.5**	**56.7**	**66.0**	**61.3**
TGS-Agent	36.4	47.6	42.0	44.9	53.8	49.3
+ MQ-Auditor	**37.1**	**49.6**	**43.4**	**46.7**	**55.7**	**51.2**

不同的交并比估计方式。 MQ-Auditor 通过下一词元预测的方式生成交并比值，直接估计交并比，这种方式称为“直接NTP”。或者，也可以将交并比视为由一个特殊词元表示的单个标量（例如，<交并比值>）。在这种情况下，MQ-Auditor 首先生成该特殊词元，然后将最终大语言模型解码器层中对应的嵌入输入一个带有 Sigmoid 激活函数的线性回归头，这种方式称为“回归ST”。回归得到的交并比与真实标注之间的均方误差会与下一词元预测损失结合用于训练。如表 5所示，“回归ST”策略对全负掩码类型（交并比 =0）尤其有效，均方根误差显著更低。然而，它在其他掩码类型上的表现明显不如“直接NTP”。这可能是因为这些掩码类型的交并比值在 $(0,\ 1)$范围内变化较大，难以通过单个线性层准确回归耳层。

训练集正样本比例 p。如第4节（“训练集”）所述，我们将正样本比例定义为 p 每个小批次中正样本（即具有完美掩码的实例）的占比。表6给出了不同 p取值下的消融结果。较高的比例（$p = 75\%$）可降低均方根误差，并提升 F_2 分数在完美类型上的表现，但会导致负掩码类型的性能大幅下降。相反，当 $p = 25\%$时，模型在大多数负掩码类型上的 F_2 分数更高，而均方根误差仍相对较高。总体而言，$p = 50\%$在所有掩码类型上的交并比估计准确率与分类性能之间实现了最佳权衡。

5.3. 通过 MQ-Auditor 改进分割

第 5.1 节的结果表明，MQ-Auditor 在评估候选掩码质量方面的有效性和优越性。值得注意的是，在推理过程中，MQ-Auditor 无需访问真值掩码，即可估计交并比或预测掩码类型及建议操作。这一特性使 MQ-Auditor 能够在实际场景中无缝集成到现有的 Ref-AVS 分割模型中。具体而言，我们在两种先前最先进的方法 EEMC [54] 和 TGS-Agent [79]上验证了这一能力。我们首先使用这些方法生成分割掩码，然后由 MQ-Auditor 对其进行评估。我们找出那些掩码被 MQ-Auditor 预测为全负类型的视频样本；这种类型代表最常见的失败模式，

即分割模型错误地分割了非目标对象。由于 MQ-Auditor 还会在推理输出中给出目标对象类别，我们利用这一信息提示 Grounded-SAM2 [48]重新生成分割掩码。我们报告了应用基于 MQ-Auditor 的质量评估前后的分割性能。如表 7所示，MQ-Auditor 始终能够提升两种基线方法的性能。具体而言，对于 EEMC 基线，在未见测试样本上的 Jaccard指数 $\mathcal{J}$和 F 分数 $\mathcal{F}$ 分别提升了 40% 和 33.5%。在这种设置下，MQ-Auditor 充当一个反思智能体，能够自动诊断掩码质量并提供可执行的指导，包括修正后的目标对象线索，以支持掩码修订。定性示例如图 ***11*** 和 ***12***所示。我们相信，随着未来出现更强大的掩码质量评估模型，这类改进还能进一步提升。

6. 结论

本文研究了语言指代的音视分割中一个长期被忽视的问题——掩码质量评估（MQA），即在测试时无法获取真实标注的情况下评估分割掩码。我们构建了 MQ-RAVSBench 数据集，并提出了一种基线方法 MQ-Auditor。大量实验表明，MQ-Auditor 的表现优于强大的开源及商业多模态大语言模型，并能有效补充现有的 Ref-AVS 系统。我们希望本研究能推动未来工作不再局限于掩码生成，而是转向配备显式错误诊断、质量审计和自动纠正机制的分割系统。

更广泛的影响。 本研究旨在通过实现分割掩码的自动质量评估，提升多模态分割系统的可靠性和可解释性。所提出的掩码质量评估框架 MQ-Auditor 能够提供结构化且可执行的反馈，有助于减少下游分割应用中的静默失败。然而，与其他基于多模态大语言模型的模型一样，MQ-Auditor 可能会继承其训练数据和底层大语言模型中的偏差，从而影响其在不同对象类别或环境下错误评估的一致性。因此，在安全关键场景中部署此类系统时必须谨慎，因为人工监督仍然至关重要。

参考文献

[1] Inclusion AI、Bowen Ma、Cheng Zou、Canxiang Yan、Chunxiang Jin、Chunjie Shen、Chenyu Lian、Dandan Zheng、Fudong Wang、Furong Xu等。Ming-Flash-Omni：一种用于多模态感知与生成的稀疏统一架构。*arXiv*预印本 *arXiv:2510.24821*，2025年。6，7，16，17

[2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025。3，18[3] Zhuoxuan Cai、Jian Zhang、Xinbin Yuan、Peng-Tao Jiang、Wenxiang Chen、Bowen Tang、Lujian Yao、Qiyuan Wang、Jin wen Chen 和 Bo Li。Q-ponder：基于推理的视觉质量评估统一训练流程。arXiv 预印本 arXiv:2506.05384，2025。13[4] Linhan Cao、Wei Sun、Weixia Zhang、Xiangyang Zhu、Jun Jia、Kaiwei Zhang、Dandan Zhu、Guangtao Zhai 和 Xiongkuo Min。Vqathinker：通过强化学习探索可泛化且可解释的视频质量评估。arXiv 预印本 arXiv:2508.06051，2025。13[5] Sanyuan Chen、Yu Wu、Chengyi Wang、Shujie Liu、Daniel Tompkins、Zhuo Chen 和 Furu Wei。BEATs：使用声学分词器进行音频预训练。arXiv 预印本 arXiv:2212.09058，2022。5[6] Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qinglong Zhang、Xizhou Zhu、Lewei Lu 等。Internvl：扩展视觉基础模型并使其适用于通用视觉语言任务。在 CVPR，页 24185–24198，2024。13[7] Bowen Cheng、Anwesa Choudhuri、Ishan Misra、Alexander Kirillov、Rohit Girdhar 和 Alexander G Schwing。用于视频实例分割的 Mask2Former。arXiv 预印本 arXiv:2112.10764，2021。13[8] Haoyue Cheng、Zhaoyang Liu、Hang Zhou、Chen Qian、Wayne Wu 和 Limin Wang。弱监督音视频视频解析的联合模态标签去噪。在 ECCV，页 431–448，2022。13[9] Zesen Cheng、Sicong Leng、Hang Zhang、Yifei Xin、Xin Li、Guanzheng Chen、Yongxin Zhu、Wenqi Zhang、Ziyang Luo、Deli Zhao 等。Videollama 2：推进视频大语言模型中的时空建模与音频理解。arXiv 预印本 arXiv:2406.07476，2024。13[10] Henghui Ding、Song Tang、Shuting He、Chang Liu、Zuxuan Wu 和 Yu-Gang Jiang。多模态指称分割：综述。arXiv 预印本 arXiv:2508.00265，2025。1[11] Henghui Du、Guangyao Li、Chang Zhou、Chunjie Zhang、Alan Zhao 和 Di Hu。Crab：一种通过显式协作实现统一音视频场景理解的模型。在 CVPR，页 18804–18814，2025。13[12] Huiyu Duan、Qiang Hu、Jiarui Wang、Liu Yang、Zitong Xu、Lu Liu、Xiongkuo Min、Chunlei Cai、Tianxiao Ye、Xiaoyun Zhang 等。FineVQ：细粒度用户生成内容视频质量评估。在 CVPR，页 3206–3217，2025。13[13] Junyu Gao、Mengyuan Chen 和 Changsheng Xu。为弱监督音视频事件感知学习概率性存在-缺失证据。IEEE TPAMI，2025。13[14] Shengyi Gao、Zhe Chen、Guo Chen、Wenhai Wang 和 Tong Lu。Avsegformer：基于 Transformer 的音视频分割。在 AAAI，第 38 卷，页 12155–12163，2024。13[15] Qihang Ge、Wei Sun、Yu Zhang、Yunhao Li、Zhongpeng Ji、Fengyu Sun、Shangling Jui、Xiongkuo Min 和 Guangtao

Zhai。Lmm-vqa：利用大型多模态模型推进视频质量评估。IEEE TCSVT，2025。13[16] Tiantian Geng、Teng Wang、Jinming Duan、Runmin Cong 和 Feng Zheng。在未剪辑视频中密集定位音视频事件：大规模基准与基线。在 CVPR，页 22942–22951，2023。13 – 。[17] 谷歌。Gemini：一系列功能强大的多模态模型。https://deepmind.google/models/gemini/pro/，2025。2，6，7，17[18] Cunhan Guo、Heyan Huang 和 Yanghao Zhou。利用分层编码器和音频引导增强音视频分割。Neurocomputing，594：127885，2024。13[19] Ruohao Guo、Xianghua Ying、Yaru Chen、Dantong Niu、Guangyao Li、Liao Qu、Yanyu Qi、Jinxing Zhou、Bowei Xing、Wenzhen Yue 等。音视频实例分割。在 CVPR，页 13550–13560，2025。13[20] Shaofei Huang 、 Rui Ling 、Tianrui Hui、Hongyu Li、Xu Zhou、Shifeng Zhang、Si Liu、Richang Hong 和 Meng Wang。重新审视以视觉为中心的 Transformer 音视频分割。在 CVPR，页 8352–8361，2025。13[21]Qing Jiang、Junan Huo、Xingyu Chen、Yuda Xiong、Zhaoyang Zeng、Yihao Chen、Tianhe Ren、Junzhi Yu 和 Lei Zhang。通过下一点预测检测任意目标。arXiv 预印本 arXiv:2510.12798，2025。4[22] Dian Jin、Yanghao Zhou、Jinxing Zhou、Jiaqi Ma、Ruohao Guo 和 Dan Guo。Simtoken：用于音视频指称分割的简单基线。在 ICASSP，页 1–5，2026。13[23] Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milanfar 和 Feng Yang。Musiq：多尺度图像质量变换器。在 ICCV，页 5148–5157，2021。13[24] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。万物分割。在 ICCV，页 4015–4026，2023。13[25] Yevheniia Kryklyvets、Mohammed Irfan Kurpath、Sahal Shaji Mullappilly、Jinxing Zhou、Fahad Shahbaz Khan、Rao Muhammad Anwer、Salman Khan 和 Hisham Cholakkal。Mavis：面向鸟类物种的多模态对话助手。在 EMNLP，页 28601–28627，2025。13[26] Mohammed Irfan Kurpath、Jaseel Muhammad Kaithakkodan、Jinxing Zhou、Sahal Shaji Mullappilly、Mohammad Almansoori、Noor Ahsan、Beknur Kalmakhanbet、Sambal Shikhar、Rishabh Lalla、Jean Lahoud 等。长视频全模态推理与工具使用的基准及智能体框架。arXiv 预印本 arXiv:2512.16978，2025。13[27] Yung-Hsuan Lai、Yen-Chun Chen 和 Frank Wang Yu-Chiang。模态无关教师与弱监督音视频事件解析器相遇。在 NeurIPS，页 1–19，2023。13[28] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。Blip-2：利用冻结图像编码器和大语言模型实现语言-图像预训练的引导。在 ICML，页 19730–19742，2023。5 – 。[29] Jia Li 和 Yapeng Tian。从波形到像素：音视频分割综述。arXiv 预印本 arXiv:2508.03724，2025。1[30] Kexin Li、Zongxin Yang、Lei Chen、Yi Yang 和 Jun Xiao。Catr：用于音视频视频分割的组合依赖音频查询 Transformer。收录于 ACM MM，第 1485–1494 页，2023。13[31] Weiqi Li、Xuanyu Zhang、Shijie Zhao、Yabin Zhang、Junlin Li、Li Zhang 和 Jian Zhang。Q-insight：通过视觉强化学习理解图像质量。arXiv 预印本 arXiv:2503.22679，2025。13[32] Zhangbin Li、Dan Guo、Jinxing Zhou、Jing Zhang 和 Meng Wang。用于音视频问答的物体感知自适应正性学习。收录于 AAAI，第 3306–3314 页，2024。13[33] Zhangbin Li、Jinxing Zhou、Jing Zhang、Shengeng Tang、Kun Li 和 Dan Guo。用于音视频问答的补丁级发声物体跟踪。收录于 AAAI，第 5075–5083 页，2025。13[34] Yan-Bo Lin、Yu-Jhe Li 和 Yu-Chiang Frank Wang。用于音视频事件定位的双模态 seq2seq 网络。收录于 ICASSP，第 2002–2006 页，2019年。13[35] Chen Liu、Liying Yang、Peike Li、Dadong Wang、Lincheng Li 和 Xin Yu。动态推导与消除：通过增强音频语义实现音频视觉分割。收录于 CVPR，第 3131–3141 页，2025。13[36] Wuyang Liu、Yi Chai、Yongpeng Yan 和 Yanzhen Ren。竖屏短视频中的音视频事件定位。arXiv 预印本 arXiv:2504.06884，2025。13[37] Xu Liu、Na Xia、Jinxing Zhou、Zhangbin Li 和 Dan Guo。通过多模态交互式脉冲神经网络实现高能效音视频分类。ACM TOMM，第 1–24 页，2025。13[38] Ziyang Luo、Nian Liu、Fahad Shahbaz Khan 和 Junwei Han。Aurora：通过结构化推理和强化学习增强指代音视频分割理解。收录于 AAAI，第 1–17 页，2026。13[39]Ziyang Luo、Nian Liu、Xuguang Yang、Salman Khan、Rao Muhammad Anwer、Hisham Cholakkal、Fahad Shahbaz Khan 和 Junwei Han。Tavis：基于基础模型的文本桥接音视频分割。收录于 ICCV，第 1–10 页，2025。13 –。[40] Juncheng Ma、Peiwen Sun、Yaoting Wang 和 Di Hu。垫脚石：音视频语义分割的渐进式训练策略。收录于 ECCV，第 311–327 页。Springer，2024年。13[41] Yuxin Mao、Xuyang Shen、Jing Zhang、Zhen Qin、Jinxing Zhou、Mochu Xiang、Yiran Zhong 和 Yuchao Dai。Tavg-bench：文本到可听视频生成基准测试。收录于 ACM MM，第 6607–6616 页，2024年。13[42] Yuxin Mao、Jing Zhang、Mochu Xiang、Yiran Zhong 和 Yuchao Dai。基于多模态变分自编码器的音视频分割。收录于 ICCV，第 954–965 页，2023。13[43] Shervin Minaee、Yuri Boykov、Fatih Porikli、Antonio Plaza、Nasser Kehtarnavaz 和 Demetri Terzopoulos。基于深度学习的图像分割：综述。IEEE TPAMI，44(7):3523–3542，2021。1[44]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry，

Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。收录于 ICML，第 8748–8763 页，2021年。5[45] Abduljalil Radman 和 Jorma Laaksonen。Tsam：以多模态提示增强 SAM 的时间建模，用于指代表达音视频分割。收录于 CVPR，第 23947–23956 页，2025。13 –。[46] Varshanth Rao、Md Ibrahim Khalil、Haoda Li、Peng Dai 和 Juwei Lu。用于音视频事件定位的双视角网络。收录于 ECCV，第 689–704 页，2022年。13[47] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman R¨adle、Chloe Rolland、Laura Gustafson、Eric Mintun、Junt- ing Pan、Kalyan Vasudev Alwala、Nicolas Carion、Chao- Yuan Wu、Ross Girshick、Piotr Doll´ar 和 Christoph Feicht- enhofer。SAM 2：分割图像和视频中的万物。arXiv 预印本 arXiv:2408.00714，2024年。4，13[48] Tianhe Ren、Shilong Liu、Ailing Zeng、Jing Lin、Kunchang Li、He Cao、Jiayu Chen、Xinyu Huang、Yukang Chen、Feng Yan 等。Grounded SAM 2：在视频中定位并跟踪万物。https://github.com/IDEA-Research/Grounded-SAM-2，2024年。9，17[49] Jinbae Seo、Hyeongjun Kwon、Kwonyoung Kim、Jiyoung Lee 和 Kwanghoon Sohn。学习听辨：增强声源关联，以实现稳健的视听实例分割。arXiv 预印本 arXiv:2509.22740，2025。13[50] Xuyang Shen、Dong Li、Jinxing Zhou、Zhen Qin、Bowen He、Xiaodong Han、Aixuan Li、Yuchao Dai、Lingpeng Kong、Meng Wang 等。细粒度可听视频描述。收录于 CVPR，第 10585–10596 页，2023。13[51] Yapeng Tian、Dingzeyu Li 和 Chenliang Xu。统一的多感官感知：弱监督音视视频解析。收录于 ECCV，第 436–454 页，2020年。13[52] Yapeng Tian、Jing Shi、Bochen Li、Zhiyao Duan 和 Chenliang Xu。无约束视频中的音视频事件定位。收录于 ECCV，第 247–263 页，2018年。13[53] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等。Llama 2：开放基础模型和微调聊天模型。arXiv 预印本 arXiv:2307.09288，2023。6[54] Yaoting Wang、Peiwen Sun、Dongzhan Zhou、Guangyao Li、Honggang Zhang 和 Di Hu。Ref-AVS：在音视场景中指代并分割物体。收录于 ECCV，第 196–213 页。Springer，2024年。1，2，3，8，13，17，28[55] Yuji Wang、Haoran Xu、Yong Liu、Jiaze Li 和 Yansong Tang。SAM2-LOVE：语言辅助音视场景中的万物分割模型 2。收录于 CVPR，第 28932–28941 页，2025。13 –。[56] Haoning Wu、Zicheng Zhang、Erli Zhang、Chaofeng Chen、Liang Liao、Annan Wang、Kaixin Xu、Chunyi Li、Jingwen Hou、Guangtao Zhai 等。Q-Instruct：提升多模态基础模型的低级视觉能力。收录于 CVPR，第 25490–25500 页，2024年。13[57] Haoning Wu、Zicheng Zhang、Weixia Zhang、Chaofeng Chen、Liang Liao、Chunyi Li、Yixuan Gao、Annan Wang、Erli Zhang、Wenxiu Sun 等。Q-Align：教会大型多模态模型进行基于离散文本定义等级的视觉评分。arXiv预印本 arXiv:2312.17090，2023。13[58] Yu Wu 和 Yi Yang。探索弱监督音视频视频解析中的异构线索。载于 CVPR，第 1326–1335 页，2021年。13[59] Yan Xia 和 Zhou Zhao。用于音视频事件定位的跨模态背景抑制。载于 CVPR，第 19989–19998 页，2022年。13[60] Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等。Qwen2. 5-omni 技术报告。arXiv预印本 arXiv:2503.20215，2025年。6, 7, 17[61] Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等。Qwen2. 5-omni 技术报告。arXiv预印本 arXiv:2503.20215，2025年。16[62] Kangmin Xu、Liang Liao、Jing Xiao、Chaofeng Chen、Haoning Wu、Qiong Yan 和 Weisi Lin。通过高效的 Transformer 适配与局部特征增强提升图像质量评估。载于 CVPR，第 2662–2672 页，2024年。13[63] Junfeng Yang、Jing Fu、Zhen Zhang、Limei Liu、Qin Li、Wei Zhang 和 Wenzhi Cao。Align-iqa：通过可定制指导，使图像质量评估模型与多样化的人类偏好相一致。载于 ACM MM，第 10008 10017 页，2024 13 – 。[64] Kaining Ying、Henghui Ding、Guangquan Jie 和 Yu-Gang Jiang。迈向指代表达音视分割中的全模态表达与推理。载于 ICCV，第 22575– 22585 页，2025年。13[65] Zhiyuan You、Xin Cai、Jinjin Gu、Tianfan Xue 和 Chao Dong。教大语言模型利用分数分布回归出准确的图像质量分数。载于 CVPR，第 14483–14494 页，2025年。13[66] Zhiyuan You、Zheyuan Li、Jinjin Gu、Zhenfei Yin、Tianfan Xue 和 Chao Dong。超越分数的描述：通过多模态语言模型推进图像质量评估。载于 ECCV，第 259–276 页，2024年。13[67] Xiao Yu、Yan Fang、Xiaojie Jin、Yao Zhao 和 Yunchao Wei。Prefm：通过预测未来建模实现在线音视频事件解析。arXiv预印本 arXiv:2505.23155，2025年。13[68] Boqiang Zhang、Kehan Li、Zesen Cheng、Zhiqiang Hu、Yuqian Yuan、Guanzheng Chen、Sicong Leng、Yuming Jiang、Hang Zhang、Xin Li 等。Videollama 3：用于图像和视频理解的前沿多模态基础模型。arXiv预印本 arXiv:2501.13106，2025年。6, 7, 16, 17[69] Zicheng Zhang、Ziheng Jia、Haoning Wu、Chunyi Li、Zi-jian Chen、Yingjie Zhou、Wei Sun、Xiaohong Liu、Xiongkuo Min、Weisi Lin 等。Q-bench-video：基准测试大型多模态模型的视频质量理解能力。载于 CVPR，第 3229 3239 页，2025 13 – 。[70] Pengcheng Zhao、Jinxing Zhou、Yang Zhao、Dan Guo 和 Yanxiang Chen。用于音视频视频解析的多模态类别感知语义增强网络。载于 AAAI，第 10448–10456 页，2025年。13[71] Jinxing Zhou、Dan Guo、Ruohao Guo、Yuxin Mao、Jingjing Hu、Yiran Zhong、Xiaojun Chang 和 Meng Wang。迈向

开放词汇音视频事件定位。载于 CVPR，第 8362–8371 页，2025年。13[72] Jinxing Zhou、Dan Guo、Yuxin Mao、Yiran Zhong、Xiaojun Chang 和 Meng Wang。用于音视频视频解析的标签预判事件解耦。载于 ECCV，第 35– 51 页，2024年。13[73] Jinxing Zhou、Dan Guo 和 Meng Wang。沿音视频事件线传播对比正样本。IEEE TPAMI，45(6):7239–7257，2022年。13[74] Jinxing Zhou、Dan Guo、Yiran Zhong 和 Meng Wang。通过分段级伪标签推进弱监督音视频视频解析。IJCV，132 5329，2024 11 5308 13 ( ): – 。[75] Jinxing Zhou、Zhihui Li、Yongqiang Yu、Yanghao Zhou、Ruohao Guo、Guangyao Li、Yuxin Mao、Mingfei Han、Xiaojun Chang 和 Meng Wang。Mettle：面向内存高效音视频适配的元令牌学习。arXiv预印本 arXiv:2506.23271，2025年。13[76] Jinxing Zhou、Xuyang Shen、Jianyuan Wang、Jiayi Zhang、Weixuan Sun、Jing Zhang、Stan Birchfield、Dan Guo、Lingpeng Kong、Meng Wang 等。结合语义的音视频分割。arXiv预印本 arXiv:2301.13190，2023。13[77] Jinxing Zhou、Jianyuan Wang、Jiayi Zhang、Weixuan Sun、Jing Zhang、Stan Birchfield、Dan Guo、Lingpeng Kong、Meng Wang 和 Yiran Zhong。音视频分割。载于 ECCV，第 386–403 页，2022年。4, 13[78] Jinxing Zhou、Liang Zheng、Yiran Zhong、Shijie Hao 和 Meng Wang。沿音视频事件线传播正样本。载于 CVPR，第 8436–8444 页，2021年。13[79] Jinxing Zhou、Yanghao Zhou、Mingfei Han、Tong Wang、Xiaojun Chang、Hisham Cholakkal 和 Rao Muhammad Anwer。分割前先思考：用于指代表达音视分割的对象感知推理智能体。载于 AAAI，第 1–19 页，2026。2, 6, 8, 13, 17, 29[80] Jinxing Zhou、Ziheng Zhou、Yanghao Zhou、Yuxin Mao、Zhangling Duan 和 Dan Guo。Clasp：用于弱监督密集音视频事件定位的跨模态显著锚点语义传播。载于 AAAI，第 1 9 页，2026 13 – 。[81] Tianfei Zhou、Fatih Porikli、David J Crandall、Luc Van Gool 和 Wenguan Wang。深度学习视频分割技术综述。IEEE TP AMI，45(6):7099–7122，2022年。1[82] Xingyi Zhou、Rohit Girdhar、Armand Joulin、Philipp Kr¨ahenb¨uhl 和 Ishan Misra。使用图像级监督检测两万个类别。载于 ECCV，第 350– 368 页，2022年。3, 5[83] Yang-Hao Zhou、Heyan Huang、Cunhan Guo、Rong-Cheng Tu、Zeyu Xiao、Bo Wang 和 Xian-Ling Mao。Aloha：通过适配局部时空上下文增强音视频语义分割。ACM TOMM，2025年。13[84] Yang-Hao Zhou、Haitian Li、Rexar Lin、Heyan Huang、Jinxing Zhou、Changsen Yuan、Tian Lan、Ziqin Zhou、Yudong Li、Jiajun Xu、Jingyun Liao、Yi-Ming Cheng、Xuefeng Chen、Xian-Ling Mao 和 Yousheng Feng。Mtavg-bench：用于评估以多说话人对话为核心的音视频生成的综合基准测试。arXiv预印本 arXiv:2602.00607，2026。13 [85] Ziheng Zhou、Jinxing Zhou、Wei Qian、S hengeng Tang、Xi- aojun Chang 和 Dan Guo。在跨模态一致性与多时间粒度协作下进行密集音视频事件定位。发表于 *AAAI*，第 10905–10913 页，2025年13– 。

A. 相关工作

音视频场景理解 旨在分析音频和视觉信号，尤其是二者之间的对应关系与互补性，从而感知并推理动态音视场景 [25, 26, 32, 33, 41, 50, 75, 84]。该领域的早期研究主要关注 时间 理解。例如，音视频事件定位 [13, 34, 37, 46, 52, 59, 73, 78]和音视频视频解析 [8, 27, 51, 58, 70, 72, 74] 旨在沿时间轴识别音频和视觉事件。后续研究将这些任务扩展到更具挑战性的场景，包括开放词汇 [71], 在线 [67], 纵向模式 [36], 和未剪辑视频 [16, 80, 85]。近来，研究重点转向更细粒度的 空间 （以及时空）理解。一个具有代表性的方向是音视频分割，旨在以物体级别 [14, 30, 40, 42, 77], 语义级别 [18, 20, 35, 39, 76, 83], 或实例级别 [19, 49] 定位视频帧中的发声物体。与这些任务类似，Ref-AVS 同样需要对音视场景进行时空理解，下面我们将回顾与之最相关的研究。

指代表达音视分割旨在根据给定的指代表达，在有声视频中分割目标对象。早期研究 [45, 54, 55] 通常采用多模态融合策略，在提示 Mask2Former [7], 、SAM [24], 或 SAM2 [47]等分割解码器之前整合跨模态线索。例如，SAM2-LOVE [55] 提出专门的词元传播与累积机制，将视频帧中的多模态线索压缩为单个[SEG］词元，再用该词元提示 SAM2 进行分割。近来，若干方法利用强大的多模态大语言模型（MLLMs）生成融合后的 [SEG］词元 [11, 22, 38, 64],，随后由掩码解码器处理。例如，AURORA [38] 采用 Video-LLaMA2 [9], ，而 OmniAVS [64]则采用加入额外音频编码器的 InternVL2 [6]。TGS-Agent [79]不依赖 [SEG］词元，而是直接生成目标对象的显式文本描述，以指导后续的检测和分割。这些先前工作主要致力于改进分割掩码生成，而我们的工作则关注一个正交问题：掩码质量评估。这不仅需要理解 Ref-AVS 任务，还需要明确推理生成的掩码是否正确、是否可靠。此外，有效的掩码质量评估还能进一步改进分割掩码生成。

图像与视频质量评估 旨在自动预测人类对视觉内容的感知判断，通常以平均意见得分（MOS）表示。图像质量评估（IQA）关注受模糊、噪声和色彩失真等伪影影响的静态图像，而视频质量评估（VQA）则将这一问题扩展到视频，其中还必须考虑运动不一致、丢帧和时间闪烁 等时域伪影。为解决 IQA 问题，早期研究使用基于 Transformer 的骨干网络 [23, 62, 63] 或多模态大语言模型 [57, 65]直接回归 MOS。近年来，越来越多的方法将 IQA 重新定义为指令遵循与推理任务，而非单纯的分数回归问题。例如，Q-Instruct [56] 和 DepictQA [66] 构建了包含低级图像质量属性自然语言反馈的大规模数据集，并证明针对性指令微调能够将通用多模态大语言模型转化为有效的低级质量专家。随后，Q-Insight [31] 和 Q-Ponder [3] 采用强化学习，进一步提升跨领域分数准确性和自然语言质量推理能力。在 VQA 领域，Q-Bench-Video [69]和 FineVQ [12] 对多模态大语言模型的视频质量理解能力进行了基准测试，表明通用模型在处理时域伪影和细微退化方面仍面临困难。LMM-VQA [15]通过引入针对视频的时空词元化策略，使大规模视觉语言模型适用于视频。VQAThinker [4]通过结合多个针对性的奖励函数，将基于强化学习的推理扩展到 VQA，同时提升了泛化能力和可解释性。不同于评估感知 或审美 质量的 IQA 和 VQA，本文研究的 MQA 侧重于评估分割掩码相对于多模态指代输入的语义保真度。

表8. MQ-RAVSBench 中每种掩码类型的样本数量。对于每个 <视频、参考> 对，我们分别为 完美、剪切/膨胀/腐蚀 和 合并/全负 类型生成 1、2 和最多 3 个掩码。对于基于图像的 评估，掩码生成在单个关键视频帧上进行；而对于基于视频的 评估，则为全部 10 个视频帧生成掩码。

	总计	#完美	#剪切	#膨胀	#腐蚀	#合并	#全负
训练集	16,761	1,306	2,612	2,612	2,612	3,809	3,810
基于图像的评估							
测试集（已见）	5,609	437	874	874	874	1,274	1,276
测试集（未见）	3,691	303	606	606	606	785	785
基于视频的评估							
测试集（已见）	7,674	600	1,200	1,200	1,200	1,738	1,736
测试集（未见）	4,966	400	800	800	800	1,081	1,085

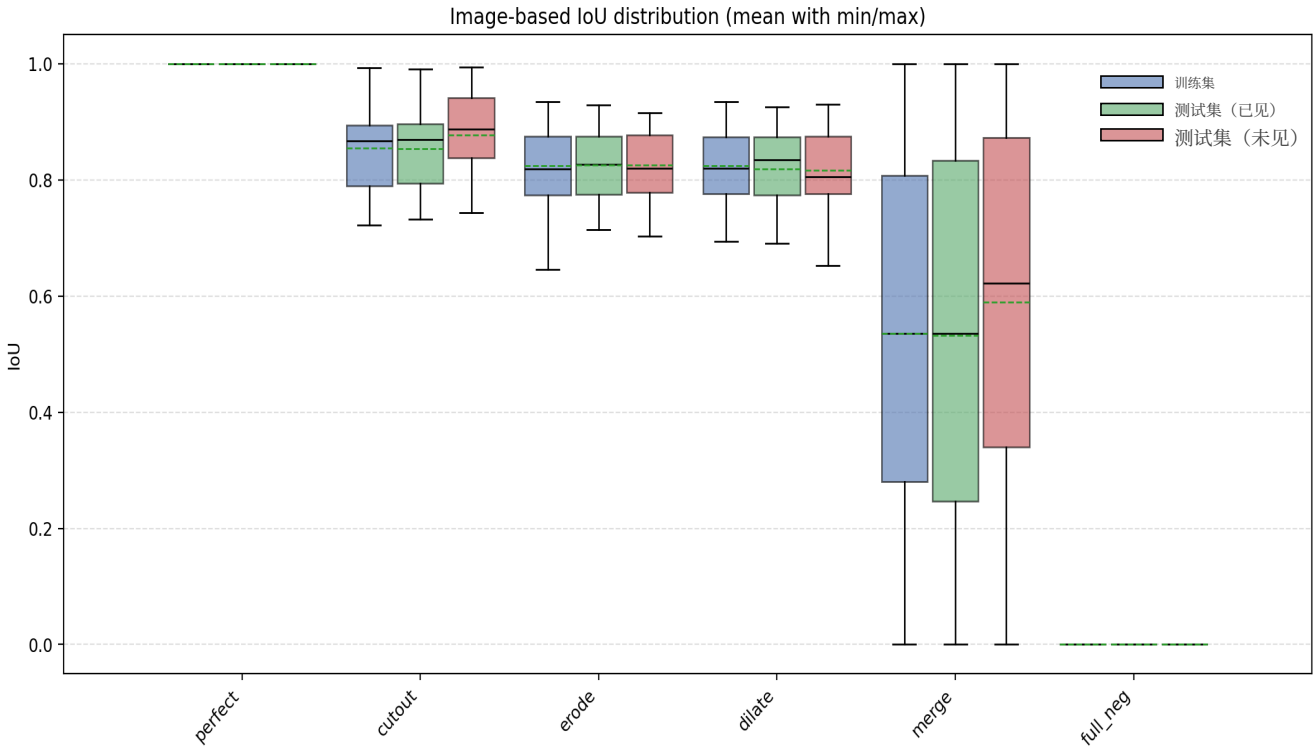


图4. MQ-RAVSBench 的 IoU 分布。对于测试集，IoU统计数据基于用于基于图像的评估的样本计算。完美和全负类型的 IoU 值始终分别为 1 和 0。剪切/膨胀/腐蚀掩码通常表现出较高的 IoU值，约为 0.8；我们有意控制这一范围，以避免过于明显的质量错误使评估变得过于简单。合并掩码的 IoU值覆盖 0 到 1 的完整范围，具体取决于真实标注对象与合并的负区域之间的相对面积。例如，当真实标注对象较小而合并的负对象较大时，生成的合并掩码会得到较低的 IoU；否则，IoU 会更高。

B. 更多数据集统计

主论文中的表 1 总结了 MQ-RAVSBench 中视频和掩码样本的总体统计信息。在表 8中，我们进一步报告了每种掩码类型的样本详细 breakdown。此外，在图 4中，我们可视化了用于 基于图像的 评估的样本的交并比分布。更多详细信息请参见相应的表格/图注。

C. 评估指标的计算细节

我们提供计算评估指标的详细公式，即均方根误差和 F_2-分数。

均方根误差。 该指标计算所有 N 次评估中预测交并比 s 与真值交并比 s_g 之间的均方根误差：

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s - s_g)^2}. \tag{8}$$

***F_2*-分数。** 掩码类型和动作预测均使用带有 $\beta = 2$ 的 F_β 分数进行评估。这种选择强调召回率，这对多模态质量评估系统尤为理想，因为：**1)** 漏掉有问题的掩码通常比错误地标记正确掩码代价更高；**2)** 我们的实证结果表明，多模态质量评估模型往往能达到较高的精确率（通常接近 100%），但召回率相对较低。F_2-分数的计算方式如下：

$$F_2 = F_\beta = (1+\beta^2)\cdot\frac{\text{P}\cdot\text{R}}{\beta^2\cdot\text{P}+\text{R}}, \quad \text{P} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{R} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{9}$$

其中，P 和 R 分别表示精确率和召回率。TP、FP 和 FN 分别表示真正例、假正例和假负例的数量。

具体而言，我们的评估遵循两种策略：

对于**基于图像的评估**，表2 中报告的已见/未见测试集最终均方根误差，是对所有 N_m 图像/掩码样本取平均后计算得到的：

$$\text{RMSE} = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_m}(s^i - s_g^i)^2}, \tag{10}$$

其中，s^i 和 s_g^i 分别表示第 i个测试样本的预测交并比和真实标注交并比。类似地，整体 F_2分数的计算方式是：首先对所有 N_m 样本累计 TP、FP 和 FN，计算每个类别的F_2，然后取宏平均：

$$F_\beta = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}F_\beta^{(c)}, \quad F_\beta^{(c)} = (1+\beta^2)\cdot\frac{\text{P}^{(c)}\cdot\text{R}^{(c)}}{\beta^2\cdot\text{P}^{(c)}+\text{R}^{(c)}}, \tag{11}$$

其中，$\mathcal{C}$ 是参与评估的类别集合（即掩码类型或操作），$\text{P}^{(c)}$和 $\text{R}^{(c)}$ 根据所有样本累计的 $\text{TP}^{(c)}$、$\text{FP}^{(c)}$和 $\text{FN}^{(c)}$ 计算得出。

对于**基于视频的评估**，表3 中报告的最终均方根误差，是先将帧级预测汇总为视频级交并比值，再对所有 N_v 测试视频取平均后计算得到的：

$$\text{RMSE} = \sqrt{\frac{1}{N_v}\sum_{i=1}^{N_v}(\overline{s}^i - \overline{s_g}^i)^2}, \quad \overline{s}^i = \frac{1}{T}\sum_{t=1}^{T}s^{i,t}, \quad \overline{s_g}^i = \frac{1}{T}\sum_{t=1}^{T}s_g^{i,t}, \tag{12}$$

其中，$\overline{s}^i$ 和 $\overline{s_g}^i$ 分别表示第 i个测试视频中 T 帧的预测交并比值和真实标注交并比值的平均值；$s^{i,t}$和$s_g^{i,t}$则分别对应第 t帧的交并比值。

类似地，视频级 F_2-分数通过汇总每个视频内的帧级真阳性、假阳性和假阴性，计算每个类别的 F_2，然后对所有视频取平均来计算：

$$F_\beta = \frac{1}{N_v}\sum_{i=1}^{N_v}F_\beta^i, \quad F_\beta^i = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}F_\beta^{i,(c)}. \tag{13}$$

这里 $F_\beta^{i,(c)}$ 是根据第 i个视频中类别 c 的$^{i,(c)}$视频级聚合真阳性、假阳性$^{i,(c)}$和假阴性$^{i,(c)}$ 计算得出的。

表 9. 掩码信息利用的消融研究。对于剪切、膨胀和腐蚀掩码类型，我们首先分别评估困难（H）和中等困难（M）样本，并报告其平均结果。“平均值”表示所有列的平均值。

指标	策略	Seen						未见						Avg.
		完美	剪切	膨胀	腐蚀	合并	全负	完美	剪切	膨胀	腐蚀	合并	全负	
均方根误差 ↓	仅掩码（$\mathcal{M}_t$）	**0.294**	0.186	**0.248**	**0.287**	0.357	0.588	0.304	0.176	0.238	0.306	0.364	0.604	0.329
	掩码帧 ($\mathcal{V}'_t$)	0.523	0.304	0.341	0.453	0.367	**0.383**	0.431	0.249	0.300	0.400	0.351	**0.407**	0.376
	掩码叠加 ($\hat{\mathcal{V}}_t$)	0.431	0.345	0.359	0.371	0.396	0.467	0.349	0.267	0.270	0.285	0.396	0.473	0.367
	两者 ($[\mathcal{M}_t;\ \mathcal{V}'_t]$)	0.359	**0.180**	0.265	0.317	**0.334**	0.507	**0.274**	**0.148**	**0.179**	**0.263**	**0.310**	0.508	**0.303**
F_2-M ↑	仅掩码 ($\mathcal{M}_t$)	**75.32**	86.25	38.44	32.57	59.63	63.84	**77.41**	**95.32**	63.00	24.46	59.98	63.28	61.62
	掩码帧 ($\mathcal{V}'_t$)	56.12	71.88	59.47	20.40	72.55	**81.78**	54.74	78.98	72.72	15.39	76.02	**81.78**	61.82
	掩码叠加 ($\hat{\mathcal{V}}_t$)	71.36	54.38	36.70	29.27	47.50	76.29	76.52	62.90	64.23	28.99	48.20	75.10	55.95
	两者（$[\mathcal{M}_t;\ \mathcal{V}'_t]$）	60.58	**87.60**	**69.15**	**41.62**	**74.55**	69.43	64.34	95.18	**82.83**	**34.28**	**77.73**	70.92	**69.02**
F_2-A ↑	仅掩码（$\mathcal{M}_t$）	**75.32**	56.46	28.13	26.65	36.36	67.46	**77.41**	**64.10**	38.73	20.33	33.39	66.30	49.22
	掩码帧（$\mathcal{V}'_t$）	56.12	48.47	42.85	18.71	39.11	**85.51**	54.74	51.81	49.20	21.69	39.27	**84.33**	49.32
	掩码叠加 ($\hat{\mathcal{V}}_t$)	71.36	39.72	28.75	29.09	34.83	79.50	76.52	40.71	45.52	**39.68**	30.63	78.86	49.60
	两者 ($[\mathcal{M}_t; \mathcal{V}'_t]$)	60.58	**57.57**	**48.16**	**32.06**	**43.00**	74.52	64.34	62.29	**59.13**	29.68	**44.57**	74.41	**54.19**

表10. 训练数据规模的消融研究。

指标	数据规模	Seen						未见						Avg.
		完美	剪切	膨胀	腐蚀	合并	全负	完美	剪切	膨胀	腐蚀	合并	全负	
均方根误差 ↓	25%	**0.053**	0.197	**0.247**	**0.253**	0.557	0.961	**0.021**	0.159	0.224	**0.222**	0.489	0.953	0.361
	50%	0.160	0.184	0.270	0.277	0.383	0.756	0.122	0.152	0.236	0.228	0.383	0.760	0.326
	100%	0.359	**0.180**	0.265	0.317	**0.334**	**0.507**	0.274	**0.148**	**0.179**	0.263	**0.310**	**0.508**	**0.303**
F_2-M ↑	25%	**97.61**	68.13	0.00	1.28	5.15	3.41	**98.14**	84.60	0.62	0.21	4.11	4.11	30.61
	50%	93.30	75.21	11.79	5.52	60.69	27.58	95.73	89.12	11.83	7.11	41.87	37.83	46.47
	100%	60.58	**87.60**	**69.15**	**41.62**	**74.55**	**69.43**	64.34	**95.18**	**82.83**	**34.28**	**77.73**	**70.92**	**69.02**
F_2-A ↑	25%	**97.61**	38.81	1.56	3.41	4.07	6.00	**98.14**	44.64	1.44	3.06	4.91	7.06	25.89
	50%	93.30	45.52	9.68	5.62	26.64	47.84	95.73	54.38	8.81	6.68	21.92	45.85	38.50
	100%	60.58	**57.57**	**48.16**	**32.06**	**43.00**	**74.52**	64.34	**62.29**	**59.13**	**29.68**	**44.57**	**74.41**	**54.19**

表11. MQ-Auditor 与 MQ-RAVSBench 上最先进的多模态大语言模型之间的效率比较。所有指标均按每个掩码样本报告。

方法	延迟（秒）↓	吞吐量（样本/秒）↑	峰值内存（GB）↓
Video-LLaMA3-7B [68]	7.1	0.141	31.1
Qwen2.5-Omni-7B [61]	295.2	0.003	74.4
Ming-Flash-Omni [1]	112.8	0.009	66.0
MQ-Auditor（我们的方法）	**4.3**	**0.233**	**15.1**

D. 更多消融结果

本节将介绍更多消融分析和效率比较。

关于掩码利用的补充研究。在第5.2节中，我们对候选掩码利用进行了消融研究，考察了原始掩码 $\mathcal{M}_t$、掩码帧 $\mathcal{V}'_t$ 及二者的组合 $[\mathcal{M}_t;\ \mathcal{V}'_t]$。在此，我们进一步研究另一种变体：使用半透明颜色（例如绿色）将掩码直接叠加到原始帧上，而不移除掩码外的区域ˆ。我们将生成的帧记为 $\mathcal{V}_t$。对比结果见表9。尽管这种基于叠加的策略（表中的“掩码叠加”）在完美和全负掩码ˆ类型上表现具有竞争力，但在其他掩码类型（包括合并）上的表现明显较差。虽然 $\mathcal{V}_t$ 突出了掩码选中的区域，但对掩码区域与非掩码区域的区分不够明确。因此，我们采用组合方式

将二值掩码 $\mathcal{M}_t$ 与掩码帧 $\mathcal{V}_t'$ 作为默认设置，因为这种方式能更好地平衡几何线索与语义线索。

训练数据规模消融实验。我们通过使用不同规模的训练视频样本训练 MQ-Auditor，探究训练数据规模的影响。如表10所示，使用较少的训练数据（例如，25%）会提高正向完美样本的比例，从而提升完美掩码类型的性能。然而，其他掩码类型的性能会显著下降，导致较低的 F_2-M 和 F_2-A 分数。以整体性能（“平均值”）为指标，随着训练数据增多，模型性能呈现明显的上升趋势。这也表明，使用较少的训练数据（例如在少样本或零样本设置下）开发模型仍非易事。

效率分析。在论文正文中，我们展示了大量实验结果，证明 MQ-Auditor 相较于最先进的多模态大语言模型的有效性和优越性。此处，我们进一步研究其推理效率。表11报告了 MQ-Auditor 与三个开源多模态大语言模型基线的每掩码样本延迟、吞吐量（样本/秒）以及 GPU 显存峰值占用。MQ-Auditor 的整体效率最佳：处理单个掩码的延迟为 4.3 秒，吞吐量为 0.233 个样本/秒，峰值显存占用为 15.1 GB。与延迟方面最强的基线 Video-LLaMA3-7B [68]（7.1 秒）相比，MQ-Auditor 的速度约为 $1.6\times$（延迟降低约 39%），吞吐量提升约 $1.65\times$，峰值显存占用也减少约 51%（15.1 GB 对比 31.1 GB）。与更大型的全模态模型相比，效率优势更加显著。这些结果进一步凸显了 MQ-Auditor 在实际系统部署中的优势。

E. 定性分析

定性比较。表2 展示了主论文中 MQ-Auditor 与多个用于掩码质量评估的先进多模态大语言模型之间的定量比较。为了提供更直观的见解，我们还在图 5∼10中加入了定性比较，其中每个示例对应一种特定的掩码类型（完美、剪切、膨胀、腐蚀、合并，以及全负）。定性观察结果与定量结果一致。Video-LLaMA3 [68]和 Qwen2.5-Omni [60] 倾向于接受大多数候选掩码。与 Video-LLaMA3 相比，Qwen2.5-Omni 的目标物体识别能力更强，但仍无法准确估计交并比或正确识别掩码类型。相比之下，Ming-Flash-Omni [1], 表现得过于保守，往往不论候选掩码的实际质量如何，都倾向于拒绝。Gemini-3-Flash [17] 在参与比较的开源多模态大语言模型中展现出更强的整体性能，并且在大多数情况下都能识别目标对象；然而，其交并比估计不如 MQ-Auditor 准确。这些定性结果进一步凸显了 MQ-Auditor 作为一种可靠的开源掩码质量评估解决方案的优势。

通过 MQ-Auditor 改进分割。 主论文中的表 7 展示了 MQ-Auditor 可以与现有的 Ref-AVS模型（包括 EEMC）集成 [54] 以及 TGS-Agent [79], , 从而提升其分割性能。我们还提供了定性示例，以说明这一实际优势。如图 11(a) 所示，目标对象是长笛，但 EEMC 错误地分割出了钢琴。MQ-Auditor 成功审核了生成的掩码，识别出错误，并提供了正确的目标对象信息。基于这一审核反馈，使用 Grounded-SAM 重新生成分割掩码2 [48], , 从而准确分割出目标对象长笛。在图 12中也可以观察到类似的改进，其中 MQ-Auditor 再次实现了有效的错误诊断和掩码细化。这些示例表明，MQ-Auditor 能够稳健地处理现有模型生成的实际分割输出，并及时识别掩码错误，以便后续修正。

失败案例分析。 尽管定量和定性结果都证明了 MQ-Auditor 的有效性，但仍然存在一些失败案例。我们讨论两个具有代表性的示例。在图13(a) 中，MQ-Auditor 正确识别出目标对象狗，但将一个腐蚀 掩码误判为完美，从而导致错误的建议操作。由于腐蚀 掩码中的所有像素都属于目标对象，此类情况尤其具有挑战性，并可能误导掩码质量评估。这一观察结果与表2 和3中的定量结果一致，结果表明腐蚀 类型比其他掩码类型更难处理。图 13(b) 展示了另一种失败场景：MQ-Auditor 能够合理评估掩码质量，但错误地理解了目标对象的语义身份。这类幻觉错误有时会在长上下文推理过程中发生。未来的工作可以引入强化学习或基于一致性的训练策略，以帮助改善物体推理与掩码质量评估之间的一致性。

F. 局限性讨论

在这项工作中，我们探索了一个新颖且实用的问题：掩码质量评估（MQA）。我们通过展示所提出的模型 MQ-Auditor 能够作为自动分割掩码质量评分器，在推理期间无需访问真值掩码即可估计交并比，证明了掩码质量评估的优势。此外，MQ-Auditor 还可以与

现有的 Ref-AVS 分割模型集成，以识别潜在错误并进一步提升整体分割性能。尽管具有这些优势，我们的工作仍存在若干局限性。正如第 E 节所述，MQ-Auditor 在某些场景下仍会出现失败案例。此外，MQ-RAVSBench 通过定义六种具有代表性的掩码类型构建而成，这些类型模拟了人工标注和模型预测中的常见错误模式，涵盖几何质量问题和语义质量问题。然而，人工标注员或真实世界模型生成的分割掩码可能比这些预定义案例复杂得多，因此难以在单个基准测试中穷尽所有可能的失败模式。尽管如此，MQ-RAVSBench 仍提供了一个可控且可衡量的测试平台，可用于初步探索掩码质量评估。我们希望这项工作能启发未来从多个方向开展研究，包括构建更多样化的数据集、通过强化学习或改进零样本泛化能力来开发审计模型，以及将掩码质量评估扩展至其他分割设置和任务。

G. 本文使用的提示词

G.1. 全负掩码构建提示词

如第3.2节所述，全负掩码的构建由 Qwen2.5-VL-72B-Instruct-AWQ [2], 指导，该模型用于生成与指代目标对象不同的负对象。下面给出详细提示词。

您是 一种先进的通用视觉分析系统，能够进行精准的目标检测和推理 g.

用户查询目标： “{目标对象}”

任务： 分析图像并检测不同的对象，严格**排除**上文提到的目标对象。按照以下具体规则对结果排序，并以 JSON 列表形式返回。

对象粒度（完整实体）：
- **仅识别完整对象：** 识别完整实体（例如“人”“猫”），而非局部部件（例如，不要输出“手”“车轮”）。
- **交互情境：** 如果有人或动物正在与目标对象互动，请识别**主体**（例如“音乐家”），而不是身体部位。

数量限制（严格）：
- **范围：** 总共输出**2 至 5 个**对象。
- **适应性：** 如果场景简单，列出 2–3 个项目；如果场景复杂，最多列出 5 个。

排序逻辑（严格优先顺序）：
1. **潜在声源：** 人类、动物、车辆、电子设备或乐器。2. **交互与情境相关性：** 与“{目标对象}”在物理或功能上相关的对象（例如，牵引绳、道路）。3. **视觉显著性：** 其他视觉上显眼的对象。 -

区分要求：
使用描述词（颜色、位置）来区分相似对象（例如，“远处的白色面包车”）。

输出格式：
仅返回一个原始 JSON 字符串列表。不要使用 Markdown 格式。

G.2. MQ-Auditor 的系统提示

在第 4 节——“网络”中，我们介绍多模态输入的嵌入方式，并提到这些输入会被发送到预定义的系统提示，如下所示：

视频:<视频开始><视频><视频结束>音频:<音频开始><音频><音频结束>给定指代表达：{参考文本}、关键帧 <图像开始><图像><图像结束> 及其分割掩码 <掩码开始><掩码><掩码结束>，请审核掩码质量。

此处，<视频开始>、<视频结束>、<音频开始>、<音频结束>、<图像开始>、<图像结束>、<掩码开始>和<掩码结束> 是固定的特殊文本标记。而 <视频>、<音频>、<图像>和 <掩码> 将被替换为真实的多模态嵌入，分别对应视频 $\mathcal{V}$、音频 $\mathcal{A}$、关键帧 $\mathcal{V}_t$，以及候选掩码 $\mathcal{M}_t$ 与掩码帧 $\mathcal{V}_t'$拼接后的嵌入。{参考文本} 也将被替换为真实的参考句子。

G.3. MQ-Auditor 的指令微调提示

图 3 展示了 MQ-Auditor 预期学习的合并掩码类型的输出。此处，我们详细列出了用于对 MQ-Auditor 进行指令微调的各类掩码类型提示。某些掩码类型可能对应多个推荐动作，并分别对应不同的提示。<审核>、</审核>、<交并比>、</交并比>、<掩码类型>、</掩码类型>、<操作>和 </操作> 是固定的特殊文本标记。{交并比值} 将替换为掩码的实际交并比值，并四舍五入至小数点后四位。{目标对象} 将替换为真实标注中的物体类别；{负对象} 则用于全负和合并掩码类型，将替换为描述负对象的简短描述性名词短语。

掩码类型：完美

建议操作：接受
审计提示： <审核> 给定掩码对 {目标对象}具有很高的保真度，轮廓描绘近乎完美。分割边界非常准确，无需进一步修改。与真实标注的交并比为 <iou>{交并比值} </iou>，其掩码类型属于 <掩码类型> 完美 </掩码类型>，建议操作为<操作>接受 </操作> </审核>

掩码类型：全负

建议操作：拒绝
审计提示： <审核> 给定掩码的重叠错误为零，但它错误地分割了一个完全不同的对象——{负对象}，而非 {目标对象}。与真实标注的交并比为 <iou> {交并比值}</iou>，其掩码类型属于 <掩码类型> 全负 </掩码类型>，建议操作为 <操作> 拒绝 </操作> </审核>

掩码类型：剪切

建议操作：小幅修订
审计提示： <审核> 给定的掩码正确勾勒出了 {TargetObject}，但存在假阴性错误和轻微内部缺陷，导致前景连通性部分丢失。其与真实标注的交并比为 <iou>{交并比值} </iou>，其掩码类型属于 <掩码类型> 剪切 </掩码类型>，建议操作为<操作>小幅修订 </操作> </审核>

建议操作：重大修订
审计提示： <审核> 给定的掩码勾勒出了 {目标对象}，但内部存在严重的假阴性错误，导致前景像素大量丢失，覆盖区域较小。与真实标注的交并比为<iou> {交并比值} </iou>，其掩码类型为 <掩码类型>剪切 </掩码类型>，建议操作为 <操作> 重大修订 </操作> </审核>

掩码类型：膨胀

建议操作： 小幅修订
审计提示： <审核> 给定的掩码正确勾勒出了 {目标对象} ，但由于边界略微扩张而产生假阳性误差，错误地包含了少量背景像素。与真实标注的交并比为<iou> {交并比值} </iou>，其掩码类型为 <掩码类型> 膨胀 </mask type>，建议操作为 <操作> 小幅修订 </action> </audit>

建议操作： 重大修订
审计提示： <审核> 给定的掩码勾勒出了 {目标对象} ，但边界严重扩张，导致大量覆盖背景区域，产生严重的假阳性误差。与真实标注的交并比为<iou> {交并比值} </iou>，其掩码类型为 <掩码类型> 膨胀 </mask type>，建议操作为 <操作> 重大修订 </action></audit>

掩码类型：腐蚀

建议操作： 小幅修订
审计提示： <audit> 给定的掩码正确勾勒出了 {目标对象}，但由于边界略微收缩而出现假阴性错误，导致边缘细节略有损失。与真实标注的交并比为 <iou> {交并比值}</iou>，其掩码类型属于 <掩码类型> 腐蚀 </mask type>，建议操作为 < 操作 > 小幅修订 </action> </audit>

建议操作： 重大修订
审计提示： <audit> 给定的掩码勾勒出了 {目标对象}，但存在严重的边界收缩，导致严重的假阴性错误，最终形状极不准确。与真实标注的交并比为 <iou>{交并比值} </iou>，其掩码类型属于 <掩码类型> 腐蚀 </mask type>，建议操作为<操作>重大修订 </action> </audit>

掩码类型：合并

建议操作： 小幅修订
审计提示： <audit> 给定的掩码正确勾勒出了 {目标对象} ，但包含了部分 {负对象}，因此存在局部错误。与真实标注的交并比为 <iou> {交并比值} </iou>，其掩码类型属于 <掩码类型> 合并 </mask type>，建议操作为 <操作> 小幅修订 </action> </audit>

建议操作： 重大修订
审核提示： <审核> 给定的掩码正确捕获了 {目标对象} ，但也错误地合并了 {负对象}的很大一部分。与真实标注的交并比为 <交并比> {交并比数值} </交并比>，其掩码类型属于 <掩码类型> 合并 </掩码类型>，建议操作为 <操作>重大修订 </操作> </审核>

建议操作： 拒绝
审核提示： <审核> 给定的掩码存在严重错误，在分割 {目标对象} 的同时，还合并了过大面积的非目标 {负对象}，导致交并比相当低。与真实标注的交并比为 <交并比> {交并比数值}</交并比>，其掩码类型属于 <掩码类型> 合并 </掩码类型>，建议操作为 <操作>拒绝</操作> </审核>

G.4. 用于评估其他多模态大语言模型的提示

表2 对比了我们的 MQ-Auditor 与多个强大的开源和闭源多模态大语言模型。我们列出了用于评估这些多模态大语言模型的详细提示。

文本查询（要分割的目标对象）：{
ref参考文本 }
给定视频、音频、目标对象出现的参考视频帧，以及该参考帧的候选二值分割掩码，你的任务是评估候选掩码对指定目标对象的质量，并且恰好输出一个 <审核> ... </审核> 块。

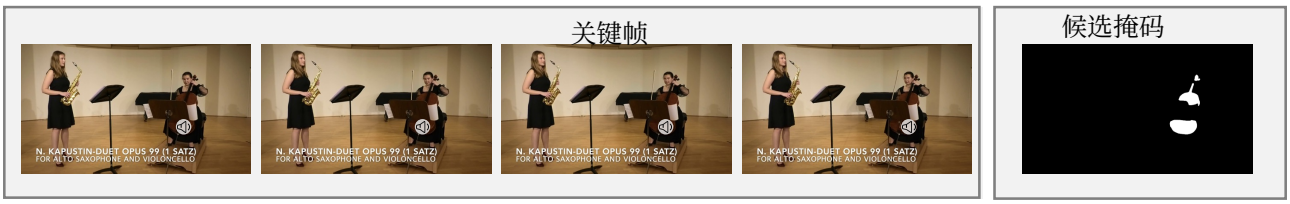
音频： {真实音频数据}**视频：** {真实视频数据}**关键视频帧：** {真实关键帧}**二值掩码：** {真实二值掩码}

你必须判断掩码是否紧密且准确地覆盖了文本中提到的对象；如有帮助，可结合视觉和音频线索进行推理。

共有 6 种可能的掩码类型（必须且只能使用以下一个关键词）：**完美：** 掩码与真实标注目标对象完美匹配。交并比=1，操作：接受**剪切/膨胀/腐蚀：** 掩码覆盖了正确的目标对象，但存在缺失、外扩或收缩的区域。交并比在 [0.85, 0.90) 之间，操作：小幅修订。交并比在 [0.75, 0.80) 之间，操作：重大修订。**合并：** 掩码包含正确的对象，但也合并了来自无关对象的额外区域。- 小范围合并：交并比在 [0.90, 1.0) 之间，操作：小幅修订。- 中等范围合并：交并比在 [0.75, 0.90) 之间，操作：重大修订。- 大范围合并：交并比在 [0.0, 0.75) 之间，操作：拒绝。**全负：** 掩码完全不包含目标对象的任何像素。交并比 = 0.0。操作：拒绝。

-

你必须输出一个单独的 <审核> 块，结构如下：<审核>[掩码质量推理] 与真实标注的交并比为 <iou> 交并比 </iou>，其掩码类型属于 <mask>掩码类型 </mask>，建议操作为 <action> 操作 </action>。</audit>



参考：持续发声时间最长的物体。

目标对象：大提琴

真实标注交并比：1.0　　掩码类型：完美 建议操作：操作　　cept

Video-LLaMA3

<审核> 与真实标注的交并比为0.98，其掩码类型为完美，建议操作为接受。</审核>

Qwen2.5-Omni

<审核> 该掩码紧密且准确地覆盖了文本中提到的对象，并结合视觉和音频线索。该掩码完美匹配真实标注目标对象，即萨克斯管。交并比为 1，建议操作为接受。 </审核>

Ming-Flash-Omni

<审核> 该掩码未能准确覆盖作为目标对象的萨克斯管。分割似乎不完整，与对象在帧中的形状和位置不太吻合。交并比可能非常低，表明匹配效果很差。因此，掩码类型为\"完整_负，\"建议拒绝该掩码。需要大幅改进掩码，才能正确识别萨克斯管。</审核>

Gemini-3-Flash

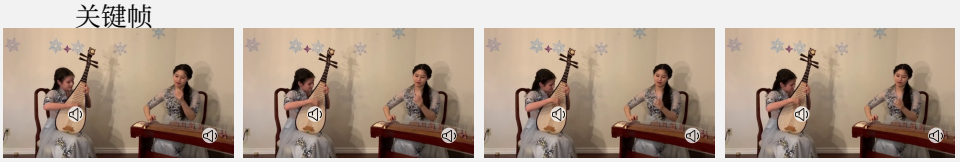
<审核> 目标对象是大提琴，它在音频中发出持续时间最长的声音。候选掩码仅覆盖大提琴的零散部分（琴身顶部和底部，以及琴头），遗漏了整个中段和琴颈。与真实标注的交并比约为0.45，其掩码类型属于剪切，建议操作为拒绝。</审核>

MQ-Auditor（我们的方法）

<审核> 给定的掩码具有很高的保真度，对大提琴的轮廓描绘近乎完美。分割边界非常准确，无需进一步修改。与真实标注的交并比为<交并比> 1。0000 </交并比>，其掩码类型属于<掩码_类型> 完美 </掩码_类型>，建议的操作是<操作> 接受 </操作> </审核>

图5. 不同掩码质量评估方法的定性比较。掩码类型：完美。

关键帧



候选掩码



参考：左侧女人演奏时发出声音的物体。

目标对象： **琵琶**

真实标注交并比：**0.8991**　　掩码类型：**剪切** 建议操作：**小幅修订**

Video-LLaMA3

<审核> Th e 与真实标注的交并比为 **1.0**，其掩码类型属于 **完美**，建议操作为 **接受**。</audi t>

Qwen2.5-Omni

<审核> 掩码紧密且准确地覆盖了文本中提到的对象，并结合了视觉和音频线索。该掩码**完美地**匹配真实标注目标对象，也就是**左侧女人正在演奏的乐器**。与真实标注的交并比为**1.0**，这意味着掩码完美匹配。因此，建议操作为<操作> **接受** </操作>。</审核>

Ming-Flash-Omni

<审核> 掩码未能准确覆盖**左侧女人正在演奏的对象**。它似乎存在错位，与乐器的形状或位置不符。掩码似乎包含了无关区域，同时遗漏了目标对象的部分区域，说明拟合效果较差。**交并比可能非常低**，表明存在明显偏差。因此，掩码类型为"**完整_负向**，"建议**拒绝**该掩码。</审核>

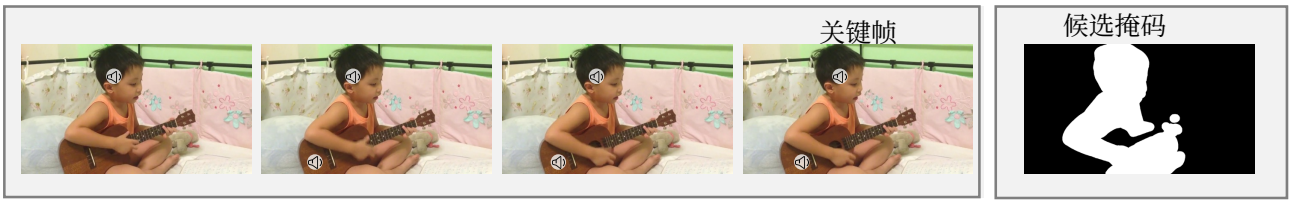
Gemini-3-Flash

<审核> 掩码正确识别了**琵琶**，即左侧女人正在演奏的乐器，但其中心区域和边缘有明显缺失。与真实标注的交并比为**0.78** IOU，其掩码类型属于<掩码> **剪切** </掩码>，建议操作为<操作> **重大修订**</操作>。</审核>

MQ-Auditor（我们的方法）

<审核> 给定的掩码正确勾勒出了**琵琶**，但存在假阴性错误和轻微的内部缺陷，导致前景连通性部分丧失。与真实标注的交并比为<交并比> **0。8910** </交并比>，其掩码类型属于<掩码_类型> **剪切** </掩码_类型>，建议采取的操作是<操作> **小幅修订** </操作></审核>

图6. 不同掩码质量评估方法的定性比较。掩码类型：剪切。



参考：通过唱歌和弹奏尤克里里发声的物体。

目标对象：**男孩**

真实标注交并比：**0.7963**　　掩码类型：**膨胀** 建议操作：**重大修订**

Video-LLaMA3

<审核> Th　与真实标注的交并比为**1.0**，其掩码类型属于**完美**，建议操作为**接受**。</audi　　t>

Qwen2.5-Omni

<审核> 掩码紧密且准确地覆盖了文本中提到的对象，同时利用了视觉和音频线索。该掩码**完美**匹配真实标注目标对象，即 **正在弹奏尤克里里并唱歌的孩子** 。与真实标注的交并比为**1.0**，建议操作为**接受**。</审核>

Ming-Flash-Omni

<审核> 掩码未能准确覆盖**正在弹奏尤克里里的孩子**，因为它看起来像是另一个对象的剪影。掩码类型为\"**完整_负**\"，因为它不包含目标对象的任何像素，建议操作为\"**拒绝**。\"</审核>

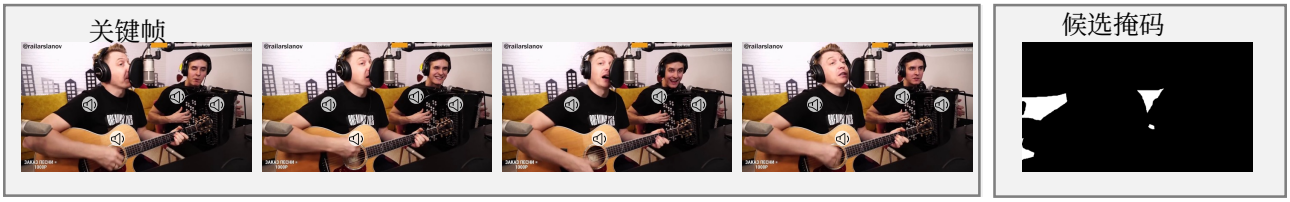
Gemini-3-Flash

<审核> 掩码正确识别了**男孩 和尤克里里** 但遗漏了男孩的双腿和尤克里里琴颈等重要部分，导致形成了剪切掩码。与真实标注的交并比为**0.78** 交并比，其掩码类型属于<掩码>**剪切** </掩码>，建议操作为<操作> **重大修订** </操作>。</审核>

MQ-Auditor（本文方法）

<审核> 给定的掩码勾勒出了**男孩**，但存在严重的边界扩张，覆盖了大面积背景，导致大量假阳性误差。与真实标注的交并比为<iou> **0。7788** </iou>，其掩码类型属于 <掩码_类型> **膨胀** </掩码_类型> ，建议采取的操作是 <操作> **重大修订**</操作> </审核>

图7. 不同掩码质量评估方法的定性比较。掩码类型：膨胀。



参考：发声的男性身后的物体。

目标对象：**沙发**

真实标注交并比： 0.7509　　掩码类型： **腐蚀** 建议操作： **重大修订**

Video-LLaMA3

<审核>与真实标注的交并比为 1.0，其掩码类型属于 完美，建议操作为 接受。</审核>

Qwen2.5-Omni

<审核> 掩码紧密且准确地覆盖了**文本中提到的对象**，同时使用了视觉和音频线索。掩码与真实标注目标对象完全匹配。交并比为1.0，这意味着掩码完美地匹配目标对象。因此，建议的操作是<操作> 接受 </操作>。</审核>

Ming-Flash-Omni

<audit>掩码完全不包含目标对象的任何像素。交并比= 0.0。操作：拒绝。\n</audit>

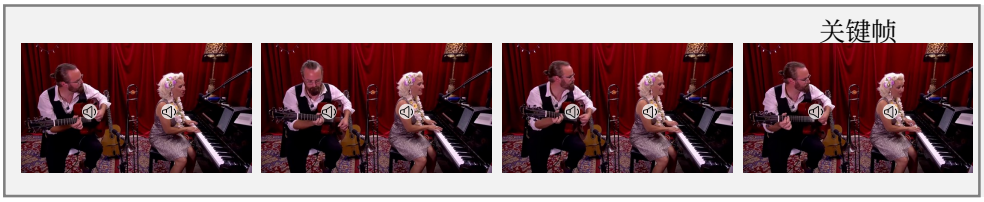
Gemini-3-Flash

<审核> 掩码识别出了**黄色沙发**在男性身后的部分区域，但极不完整，遗漏了该对象表面区域的大部分。与真实标注的交并比为<交并比> 0。4 </交并比>，其掩码类型属于<掩码> 剪切</掩码>，建议的操作是<操作> 拒绝 </操作>。</审核>

MQ-Auditor（我们的）

<审核> 给定的掩码勾勒出了**沙发**，但存在严重的边界收缩，导致显著的假阴性错误，且最终形状极不准确。与真实标注的交并比为<交并比> 0。7716 </交并比> ，其掩码类型属于<掩码_类型腐蚀 </掩码_类型>，建议的操作是<操作> **重大修订** </操作></审核>

图8。不同掩码质量评估方法的定性比较。掩码类型：腐蚀。



参考：左侧男子前方的发声物体。

目标对象：**吉他**

真实标注交并比：**0.2551**　　掩码类型：**合并** 建议操作：**拒绝**

Video-LLaMA3

<审核> I　　与真实标注的 IoU 为 **0.98**，其掩码类型为 **完美**，建议操作为 **接受**。</审核>

Qwen2.5-Omni

<审核> 掩码紧密且正确地覆盖了**文本中提到的对象**，同时利用了视觉和音频线索。掩码与真实标注目标对象完美匹配，该对象是左侧男子前方的发声物体。与真实标注的交并比为**1.0**，建议操作为**接受**。</审核>

Ming-Flash-Omni

<审核> 所提供的掩码与关键帧中的目标对象（**吉他**）不对应。它似乎是一个完全无关的掩码，可能原本用于其他对象或场景。由于掩码未包含目标对象的任何像素，交并比为**0.0**，因此应**拒绝**该掩码。\n</审核>

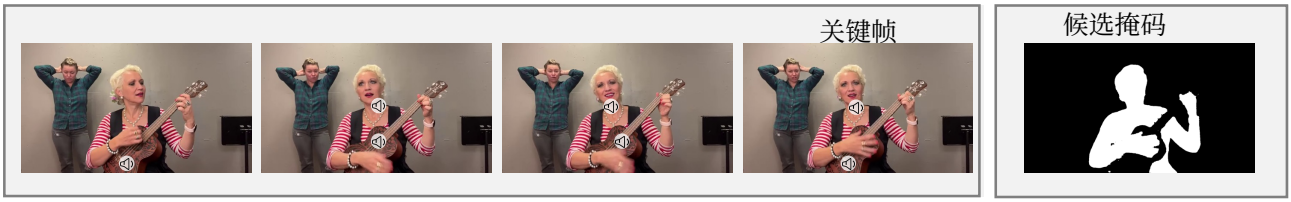
Gemini-3-Flash

<审核> 掩码**包含了目标吉他的一部分，但还合并了覆盖钢琴的大范围区域**，而钢琴是一个无关对象。与真实标注的交并比为**0.45**，其 IOU 掩码类型属于 <掩码> **合并** </掩码>，建议操作为<操作> **拒绝** </操作>。</审核>

MQ-Auditor（本文方法）

<审核> 给定的掩码存在严重错误，**在分割吉他的同时，还合并了面积过大的非目标钢琴**，因此交并比很低。与真实标注的交并比为<交并比> **0。2911** </交并比>，其掩码类型为<掩码_类型>**合并** </掩码_类型>，建议的操作为<操作> **拒绝** </操作> </审核>

图9. 不同掩码质量评估方法的定性比较。掩码类型：合并。



参考：女人手持的第一个发声物体。

目标对象：**尤克里里**

真实标注交并比：**0.0**　　掩码类型：**完整_负** 建议操作：**拒绝**

Video-LLaMA3

<审核> The I　与真实标注的 oU 为 **0.98**，其掩码类型属于 **完美**，建议操作为 **接受**。</audit>

Qwen2.5-Omni

<审核> 该掩码紧密且准确地覆盖了**文本中提到的物体**，并同时利用视觉和音频线索。该掩码与真实标注目标对象完美匹配，也就是女人手中第一个发出声音的物体。与真实标注的交并比为**1.0**，建议操作为**接受**。</审核>

Ming-Flash-Omni

<审核> 该掩码准确覆盖了女人手中的**尤克里里**，没有明显的遗漏或多余区域。分割与目标对象高度吻合，表明交并比接近**1.0**。因此，掩码类型为**完美**，建议**接受**该掩码。</审核>

Gemini-3-Flash

<审核> 目标对象是女人手中的**尤克里里**，但候选掩码**分割的是女人本身**，并明确排除了尤克里里。与真实标注的交并比为**0.0**，其掩码类型属于**完整_负类**，建议操作为**拒绝**。</审核>

MQ-Auditor（我们的方法）

<审核> 给定的掩码产生了零重叠误差，因为它错误地**将完全不同的物体——穿条纹衬衫的女人——进行了分割，而不是尤克里里**。与真实标注的交并比为<交并比> **0。0000** </交并比>，其掩码类型属于<掩码_类型> **完整_负** </掩码_类型>，建议操作为<操作> **拒绝** </操作></审核>

图10. 不同掩码质量评估方法的定性比较。掩码类型：全负。



图11。使用我们的 MQ-Auditor 前后的分割性能对比。此前最先进的 Ref-AVS 方法 EEMC [54], 用于这些示例。我们的 MQ-Auditor 可以评估真实 Ref-AVS模型生成的掩码质量，并提供正确的目标对象信息，以指导掩码细化。



图 12. 应用 MQ-Auditor 前后的分割性能对比。此前最先进的 Ref-AVS 方法 TGS-Agent [79] 用于这些示例。我们的 MQ-Auditor 能够评估真实 Ref-AVS模型生成的掩码质量，并提供正确的目标对象信息以指导掩码细化。

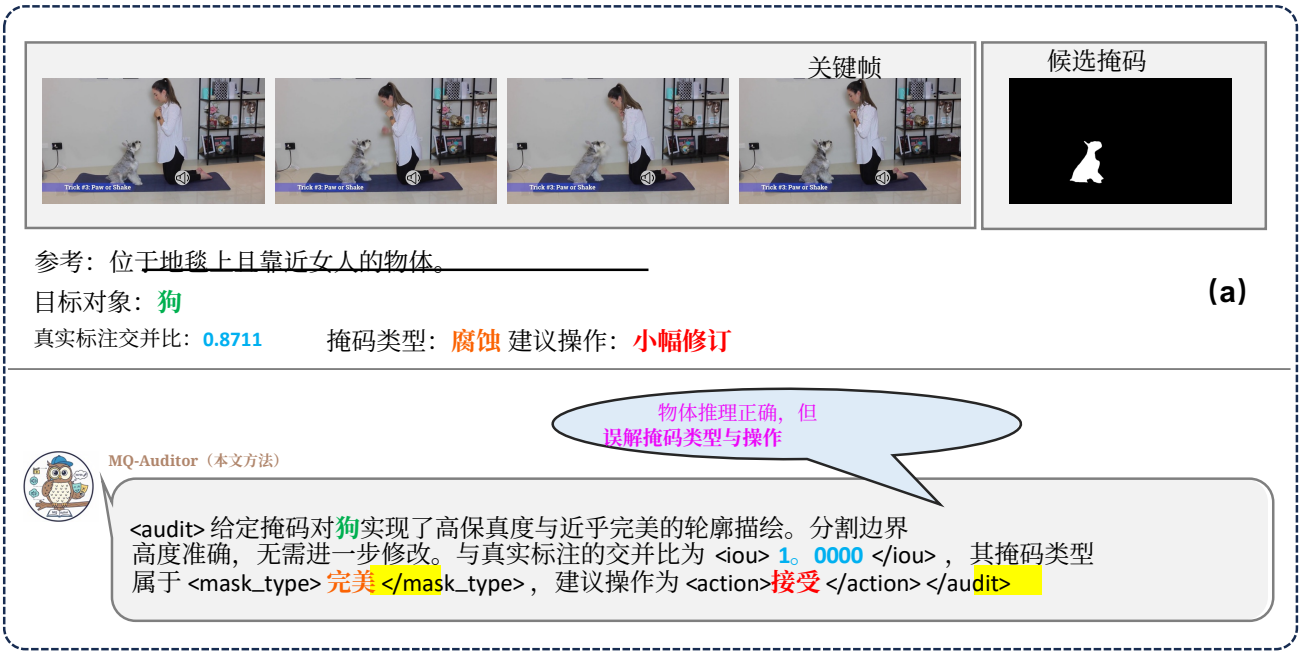


图13. 两个代表性失败案例的可视化。MQ-Auditor可能会出现掩码类型误分类或物体识别幻觉。