arXiv:2601.08871v1 [cs.SD] 12 Jan 2026

基于大型视觉语言模型的语义视觉引导声学突出

Junhua Huang、*Chao Huang*、*Chenliang Xu*

罗切斯特大学

摘要

在沉浸式叙事中，平衡对白、音乐、音效与相应视频至关重要，但当前的音频混音工作流程仍主要依靠人工，劳动强度很大。尽管近期的进展开创了视觉引导声学突出这一任务，通过多模态引导对音频源进行隐式再平衡，但哪些视觉方面能够作为有效的条件仍鲜为人知。我们通过系统研究深度视频理解能否改善音频重新混音，来填补这一空白。我们以文本描述作为视觉分析的替代方式，要求大型视觉语言模型（大型视觉语言模型）提取六类方面的信息，包括物体／角色外观、情绪、镜头焦点、色调、场景背景，以及推断出的声音相关线索。在大量实验中，镜头焦点、色调和场景背景 相较于最先进基线，始终能带来最高的感知混音质量提升。我们的研究结果（i）确定了最有力地支持连贯且与视觉一致的重新混音的视觉语义线索，并且（ii）勾勒出一条切实可行的路径，通过轻量级、由大型视觉语言模型生成的引导，实现电影级声音设计的自动化。

索引词— 视觉引导的声音突出处理、大型视觉语言模型、电影声音设计

1. 引言

电影、电视和游戏依赖视觉与声音之间紧密编排的协同配合，但如今许多音轨混音仍然依赖人工操作：行业指南估计，对白为主的内容每完成一分钟需要*1* 工程师工时，复杂项目则可能达到每完成一分钟需要 *5–10* 小时 [1]。画面锁定后的变更会触发额外的混音轮次，而视觉焦点与响度不匹配则会损害沉浸感和无障碍性。能够依照视频内容重新平衡音轨的自动化系统，有望降低成本，并让更多人享受到高质量混音。

纯音频流程（先分离，再通过启发式方法重新混音）忽略了图像，因此可能与视觉意图不一致。视觉引导声学突出（VisAH）[2]将重新混音重新定义为视觉引导的音频到音频转换，并表明，视频的文本 描述在引导效果上可媲美学习得到的视频特征。然而，VisAH 的处理粒度仅为镜头／片段级，且未能分辨哪些视觉语义最能预测理想的混音调整；特定视觉元素与混音构成之间的关系仍未得到充分探索。

为了解决这一问题，我们回顾了成熟的实践和视听领域文献：混音师通常会优先突出屏幕内声源，并通过闪避及相关控制手段保持对白清晰度 [1]。技术研究进一步表明，视觉证据有助于识别声源，并支持针对特定声源进行音量控制 [2]。这些观察表明，具体线索对于重新混音的预测作用尤其显著，例如镜头焦点、场景／背景语境、物体／角色身份、情感，以及可见的声音成因。

现代大型视觉语言模型无需笨重的视频编码器，便能以文本形式提取这类线索：它们可以定位并识别实体 [3], 与人类注视／注意力相对应 [4], , 并推理手部与物体交互 [5]。受此启发，我们研究大型视觉语言模型条件引导的重新混音：冻结的大视觉语言模型生成简洁的文本描述符，作为帧级语义的代理，我们则测试将特定子集用作条件信号的效果。我们发现，镜头焦点 、场景 和声源 是影响最大的组成部分；通过简单地改进提示词，使用大视觉语言模型提取的线索，能够取得优于纯音频和以往多模态基线的最先进效果。

2. 相关工作

从分离到重新混音。 早期的“自动混音”流程首先通过纯音频源分离估计各音轨，然后应用启发式增益曲线；常用的开源系统包括 Open-Unmix [6]和 Demucs [7]。为利用视频信息，视听对应方法将可见事件与声音对齐。与我们的设定最为接近的是《通过观看电影学习突出音频》（VisAH），该方法将视觉引导的声学突出形式化为隐式混音，而非显式分离：它通过预测单个混音掩码，将源分离与增益重新平衡统一起来，并在合成的MUDDYMIX语料库上训练端到端音频到音频框架，其表现优于纯音频基线。我们同样采用这种端到端重新混音 形式，并探究哪些视觉／文本线索最能有效引导时变增益控制。

将大型视觉语言模型作为细粒度条件信号。近期的大型

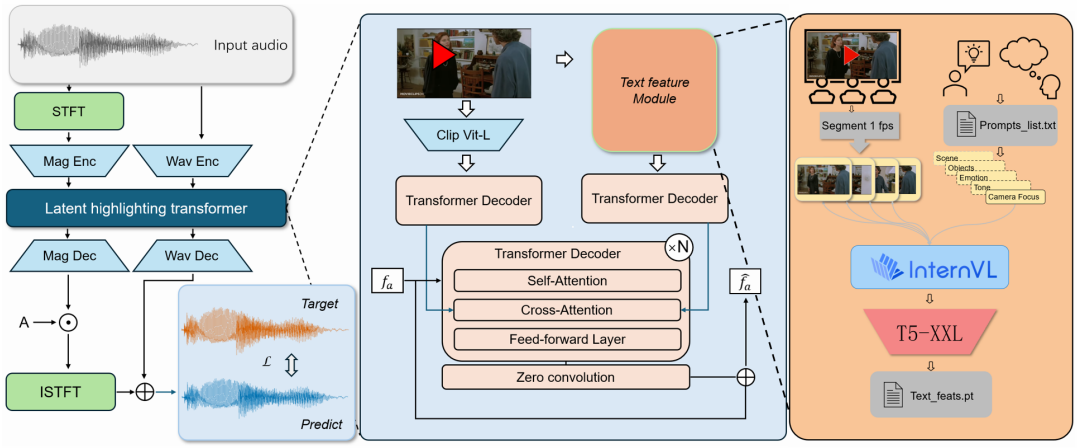

图 1。概览与 VisAH [2]相同，区别在于文本特征模块 （橙色）会馈入上下文编码器。

视觉语言模型（大型视觉语言模型）提供了与剪辑意图对齐的、基于帧的细粒度语义（例如，观众注视／注意力 [4]以及文本区域对齐 [3]）。基于这些进展，我们采用来自冻结的大视觉语言模型的文本代理（关键帧 → 字幕 → 文本嵌入），为轻量级增益控制器提供条件，并系统地消融六个语义方面：外观、情绪、色调、摄像机焦点、场景背景以及推断出的声音相关线索，以确定哪些线索最能提升混音质量。

3. 方法

3.1. 任务概述

我们研究视觉引导的声学突出处理（*VGAH*）：给定一段短视频片段 V 及其粗略或平衡不佳的^混合音轨 A，生成重新平衡后的配乐 A，在保留内容（对白、音乐、拟音）的同时，调整相对响度，使其符合画面中的叙事显著性。此前的纯音频流程和 VisAH 变体以镜头／片段级的未区分视频特征作为条件，而对哪些具体视觉语义最能预测理想的混音调整这一问题探讨不足。为此，我们提出 SemMix，研究哪些具体视觉信号（语义）能触发更好的重新平衡效果（混音）。

3.2. 流程概述

我们的架构遵循 VisAH [2]：包含双分支音频骨干网络（时域与频域）、上下文编码器、潜在突出变换器，以及结合逆短时傅里叶变换（iSTFT）的基于掩码的解码器。本文关注的是形成文本上下文的提示词路径（图 1）。

3.3. SemMix：哪些视觉语义重要?

我们的设计以既有研究的多个方向为基础。电影声音实践方面，经典文献指出，有效混音取决于一小组可感知的线索：屏幕上的情绪表演、具体的发声物体、时空场景、整体视觉氛围、可见声源以及镜头着重呈现的内容——因为这些线索会影响对白的呈现、拟音的选择、氛围营造和剪辑显著性 [8, 9]。大型视觉语言模型文献方面，研究表明，当指令含糊不清或引导模型进行画外推断时，模型会产生幻觉；提示模型仅依据可见证据作答，并在不确定时弃答，可以减轻此类错误 [10, 11]。综合这些观察，我们设计了六个主题，涵盖电影声音的关键决策要素：引导混音的故事内声音锚点、塑造氛围的场景／时间／色调线索，以及在避免背景故事或画外推断的同时，使视听注意力保持一致的焦点 [8, 12]。

- **情绪（演员）**：画面中的主要情绪；仅在次要线索清晰可见时才予以注明。
- **物体（显著）**：简要列出引人注目且与声音相关的可见的物品（名词）。
- **场景（环境与时间）**：根据画面内容推断地点、时间、光照和拥挤程度。
- **色调（色彩与情绪）**：营造氛围并影响背景音选择的色板和灯光风格。
- **声源（可见的）**：可能发出声音的屏幕内元素（场景内声源锚点）。
- **镜头焦点（显著性）**：预期观众关注的主体，以及明显表明焦点所在的线索。

3.4. 提示版本。

我们提炼出两个灵感来源不同的互补类别。聚焦型提示词会添加明确的约束，并在

表1。主要对比：最佳结果以**粗体**标出，次佳结果加下划线。指标为 $\times 100$。括号中的百分比表示相较于混合不充分的输入 的提升（越低越好）。

方法	MAG ↓	ENV ↓	KLD ↓	ΔIB ↓	W-dis ↓
混合不充分的输入	22.69	6.30	20.61	1.52	1.94
DnRv3 [13]+CDX [14]	26.32 (−16%)	7.62 (−21%)	15.87 (+23%)	1.78 (−17%)	2.84 (−46%)
Learn2Remix [15]	19.07 (+16%)	4.16 (+34%)	61.76 (−200%)	8.27（−444%）	1.20（+38%）
LCE–SepReformer [16]	17.18（+24%）	4.28（+32%）	30.99（−50%）	1.88 (−24%)	1.28 (+34%)
VisAH [2]	10.08 (+56%)	3.43 (+46%)	11.01 (+47%)	**0.80 (**+47%**)**	**0.79 (**+59%**)**
SemMix-相机聚焦	**9.99**（+56%）	**3.41**（+46%）	**10.95**（+47%）	0.87 (+43%)	**0.79**（+59%）

有帮助时提供简短的示例模式，使模型严格以可见内容为依据，并在证据不足时允许弃答 [11]。极简型提示词会尽可能缩短指令，以检验简洁的引导是否足够，并减少提示开销；这种精简做法源于相关研究发现：较长的上下文会分散模型对显著细节的注意力，并可能降低忠实性 [17]。因此，这两类提示词针对的是不同风险——规定不足与规定过度，意图并不重叠。

为通过两个示例具体说明这种对比：对于声源，在一段安静的办公室镜头中，画面里清晰可见的只有一台吊扇，聚焦版本要求始终依据屏幕内容，并在没有明确声源时返回无，从而给出单一、有依据的提示（例如风扇嗡鸣声）；最简版本则只列出可能出现在屏幕中的声源，虽然输出简短，却省略了明确的弃答规则。对于镜头焦点，在一只手握着门把手、走廊背景虚化的紧凑特写镜头中，聚焦版本会指出预期主体并说明用于实现该焦点的可见拍摄手法（例如浅景深），而最简版本只报告主体。两个版本的比较将在第5.2节中，按照与我们先前VisAH设置相同的评估指标进行。

4. 实验

4.1. 实验设置

我们沿用VisAH设置 [2]，仅修改文本路径。除非另有说明，所有数据划分、音频预处理（44.1 kHz立体声转单声道）、STFT参数、骨干网络和默认训练计划均与VisAH相同。文本嵌入的通道维度为 $C_{\text{text}} = 4096$，并通过相同的条件接口输入。模型使用Adam训练（学习率 10^{-4}），每块GPU的批量大小为12，共训练200个周期。

训练使用 $2\times$ RTX4090 GPU；完整训练约需 21 小时。

4.2. 评估指标

为将改进完全归因于提示词，我们不作任何修改地采用 VisAH 的指标套件。*MAG*↓ 量化频谱准确度，计算幅度 STFT 之间的 L1距离；而 *ENV*↓ 则通过由解析信号导出的包络，评估时域中的动态变化。在语义方面，我们计算基于 PaSST 的事件分布之间的 *KLD*↓ ，并报告基于 ImageBind 的差距 $\hat{}\Delta\text{IB} = \text{IB}(V, A) - \text{IB}(V, A)$；数值越小，表示音视频对应关系越紧密。最后，*W-dis*↓用于衡量对齐预测响度轨迹与参考响度轨迹（语音/音乐/音效）的成本，从而反映随时间变化的时序与平衡。

4.3. 基线与先前工作

我们采用 **VisAH** 使用的完整基线套件，并遵循其评估划分；我们还重新运行了已发布的配置，以确认复现结果在通常的容差范围内与原结果一致 [2]。基线如下。

•混音不佳的输入。 将数据集创建过程中人工构造的混音不佳输入作为参考。

• DnRv3 [13] + CDX [14]. 应用 DnR v3 分离器以获得三条音轨（语音、音乐、音效），然后从 CDX 分布中为每条音轨采样响度并重新混音，以生成输出。

• Learn2Remix [15] 一种端到端神经混音器；遵循 VisAH 的做法，我们将原始 Conv-TasNet 主干网络替换为 *SepReformer* ，以获得更强的分离器，并使用我们的数据进行训练。

• Listen, Chat, and Edit (LCE) [16] 一种由文本引导的混音编辑器；与 VisAH 一样，测试时无法提供明确指令，因此我们提供文本字幕作为引导。为保证公平比较，我们将 SoundEditor 主干网络替换为 *SepReformer*。

我们将 VisAH 本身视为融合原始视觉信息与文本的先前最先进水平方法；我们的贡献在于单独考察更丰富、经提示词工程设计的字幕（由 InternVL 系列字幕生成器 [18] 生成）所带来的益处，同时保持音频处理流程不变。表1总结了 MUDDYMIX 上的主要比较结果；**我们的 Sem-Mix 方法参数量减少了 18.94M，性能达到此前的最先进水平，同时模型**规模缩小了 ***9.0%***。

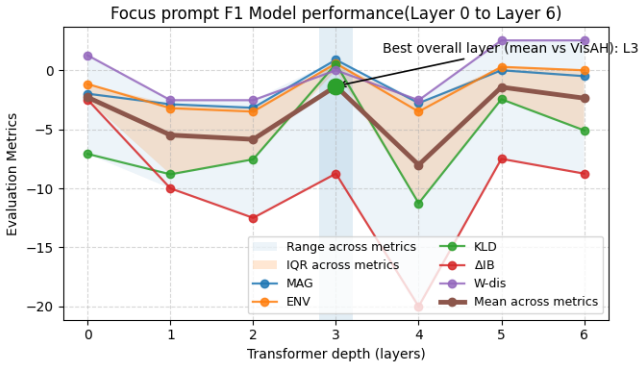

图 2。使用聚焦提示词**镜头焦点**时，模型在第 0 层至第 6 层的性能。阴影区域表示跨指标的波动范围（最小值–最大值和四分位距）。均值曲线在L=3处达到峰值；L=5/6 带来轻微的 W-dis 润色。

表 2。总体提示词比较。每列的最佳结果以**粗体**显示；优于 VisAH 的数值带下划线。改进幅度是相对于VisAH（基线）计算的。数值越低越好。

提示词	层数	MAG ↓	ENV ↓	KLD ↓
VisAH（基线）	3	10.34（0%）	3.51（0%）	11.75（0%）
情绪（演员）	3	10.32（+0.2%）	3.54（−0.9%）	11.97（−1.9%）
物体（显著）	3	10.73（−3.8%）	3.68（−4.8%）	12.07（−2.7%）
场景（场景与时间）	3	10.28 (+0.6%)	3.50 (+0.3%)	11.53 (+1.9%)
色调（色彩与情绪）	3	10.24 (+1.0%)	3.51 (0.0%)	12.10 (−3.0%)
声源（可见的）	3	10.29 (+0.5%)	3.49 (+0.6%)	11.51 (+2.0%)
镜头焦点（显著性）	3	**10.01 (+3.2%)**	**3.39 (+3.4%)**	**11.05 (+6.0%)**

5. 分析

5.1. 提示词比较。

我们在相同设置下对六种提示词设计进行了基准测试（表 2）。不同提示词的表现差异显著。**物体** 会导致所有指标变差，这表明罗列显著物品会使模型偏向视觉上引人注目、但与声学无关的细节。**场景**、**声源**和**色调**均能带来幅度（MAG）小幅但稳定的下降，而**场景**还能改善 KLD，这表明粗粒度的时空上下文有助于对齐分离目标。**镜头焦点**带来的提升最大：与基线相比，MAG 降低 3.2%，ENV 降低 3.4%，KLD 降低 6.0%；将描述重点放在镜头突出呈现的内容上，似乎能够过滤干扰项，并强调最可能主导混音的声源。**情绪**略微改善了 MAG，但使 ENV/KLD 变差，这意味着情感线索与声源结构的关联较弱。总体而言，**镜头焦点** 是表现最佳的提示词，其次是**声源**和**场景**。

表3。聚焦提示词集与精简提示词集的比较（越低越好）。请将短横线替换为测量值或相对变化量。

提示词集	MAG ↓	KLD ↓	W-dis ↓
聚焦	9.99	10.95	0.79
简约	10.29	12.36	0.78

5.2. 聚焦提示集与最简提示集。

受篇幅限制，我们对这两类提示集进行了简要比较。在相同的评估设置下（表 3），**聚焦提示集**的各项指标总体与**最简提示集**相当或略胜一筹，这表明简短但明确的线索（例如镜头焦点和场景上下文）有助于使视觉证据与混音目标对齐。这一观察与以往研究结果一致：模型输出对提示词构造和校准高度敏感 [19]，，而在目标无法直接观测的任务中，轻量的结构／引导能够提高保真度 [20]。

5.3. Transformer深度。

沿用 [2] 中探索的深度设置（0、3 和 6 层），我们在相同训练方案下使用聚焦提示词，将 Transformer 深度从 L=0 逐步调整至 L=6 。如图 2 所示，较浅的堆叠已能获得大部分收益： L=3 取得最佳 MAG、ENV 和 KLD（9.99、3.41、10.95），而 L=0的 ΔIB最低（0.82），$L\in\{5,6\}$ 则取得最佳 W-dis距离（0.77）。增加深度并不能使所有指标单调提升——值得注意的是，L=4 的 KLD 表现较差（12.25）——但随着深度从 3 增至 6，我们观察到分布对齐（W-dis）略有改善。所有深度设置均在 ∼175至198 个训练轮次内收敛，这表明一旦提示词注入显著且聚焦于帧的语义，超过 L=3层后收益便会递减。总体而言，L=3 的综合表现最佳，而 L=5/6虽然使 W-dis 略有改善，却要以 KLD 表现变差为代价。通过提示词设计注入聚焦的帧级语义，可以减少对深层自注意力的需求：三层已足以实现整体优异的质量，而进一步加深网络主要只能带来轻微的分布对齐改善。

6. 结论

以大视觉语言模型生成的文本线索作为视觉上下文，我们剥离出对重新混音至关重要的语义，并发现，镜头焦点和场景背景 最能稳定地提升感知混音的质量。与 VisAH 相比，我们的 SemMix 参数更少、采用更浅层的 Transformer，却取得了更好的效果，这表明，实现视觉对齐的音频重新混音，关键在于精准的视觉语义，而非庞大的视频编码器。

参考文献

[1] iZotope，《视频音频混音，第 2 部分：音频后期制作工作流程》，2018 年，提出了“~1 每完成一分钟耗时一小时”的经验法则。访问日期：2025-09-10。

[2] Chao Huang、Ruohan Gao、J. M. F. Tsang、Jan Kurcius、Cagdas Bilen、Chenliang Xu、Anurag Kumar 和 San-jeel Parekh，《通过观看电影学习突出音频》，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2025年6月，第23925–23935页。

[3]Zhiliang Peng、Wenhui Wang、Li Dong、Yaru Hao、ShaohanHuang、Shuming Ma 和 Furu Wei，《Kosmos-2：将多模态大语言模型与现实世界建立关联》，*arXiv*预印本 *arXiv:2306.14824*，2023年。

[4] Kun Yan、LeiJi、Zeyu Wang、Yuntao Wang、Nan Duan 和 Shuai Ma，《Voila-a：通过用户注视注意力对齐视觉语言模型》，*arXiv*预印本 *arXiv:2401.09454*，2024年，NeurIPS 2024年版本已发布。

[5] Chen Bao、Jiarui Xu、Xiaolong Wang、Abhinav Gupta 和 Homanga Bharadhwaj，《Handsonvlm：用于手部-物体交互预测的视觉语言模型》，*arXiv* 预印本 *arXiv:2412.13187*，2024年。

[6] Fabian-Robert St¨oter、Stefan Uhlich、Antoine Liutkus 和 Yuki Mitsufuji，《Open-unmix——音乐源分离的参考实现》，《开放源代码软件期刊》，第 4 卷，第 41 期，第 1667 页，2019年。

[7] AlexandreD´efossez、NicolasUsunier、L´eon Bottou 和 Francis Bach，《波形域中的音乐源分离》，*arXiv* 预印本 *arXiv:1911.13254*，2019年。

[8]Michel Chion，《视听：银幕上的声音》，哥伦比亚大学出版社，纽约，1994年。

[9] Tomlinson Holman，《电影与电视声音》，劳特利奇出版社，牛津，第3版，2010年。

[10] Anna Rohrbach、Lisa Anne Hendricks、Kaylee Burns、Trevor Darrell 和 Kate Saenko，“图像描述中的物体幻觉”，载于*2018*年自然语言处理实证方法会议论文集，比利时布鲁塞尔，2018年，第4035–4045页，计算语言学协会。

[11]Mina Huh、Fangyuan Xu、Yi-Hao Peng、Chongyan Chen、Hansika Murugu、Danna Gurari、Eunsol Choi 和 Amy Pavel，《面向盲人和低视力人群的视觉问题长篇回答》，*arXiv:2408.06303*，2024年，COLM 2024 口头报告焦点论文。

[12]Randy Thom，《为声音设计电影》，https://filmsound.org/articles/ designing_for_ sound.htm，访问日期：2025-09- 17。[13] Karn N. Watcharasupat、Chih-Wei Wu 和 Iroro Orife，《再母带处理：分离与再母带处理：一个支持多语言的电影音频源分离数据集》，发表于*2024*年 *IEEE* 第五届国际声音互联网研讨会 *(IS2)*，德国埃朗根，2024年9月。

[14] StefanUhlich、Giorgio Fabbro、Masato Hirano、Shusuke Takahashi、Gordon Wichern、Jonathan Le Roux、Dipam Chakraborty、Sharada Mohanty、Kai Li、Yi Luo、Jianwei Yu、Rongzhi Gu、Roman Solovyev、Alexander Stempkovskiy、Tatiana Habruseva、Mikhail Sukhovei 和 Yuki Mitsufuji，“2023声音解混挑战赛——电影解混赛道”，国际音乐信息检索学会汇刊（*TISMIR*），第7卷，第1期，第44–62页，2024年。

[15] Haici Yang、Shivani Firodiya、Nicholas J. Bryan 和 Minje Kim，《不要分离，学会混音：通过联合优化实现端到端神经混音》，载于*IEEE* 国际声学、语音与信号处理会议论文集 *(ICASSP)*，新加坡，2022年。

[16] Xilin Jiang、Cong Han、Yinghao Aaron Li 和 Nima Mesgarani，《聆听、聊天与编辑：通过文本引导的声景修改提升听觉体验》，2024年。

[17] Nelson F. Liu、Kevin Lin、John Hewitt、Ashwin Paranjape、Michele Bevilacqua、Fabio Petroni 和 Percy Liang，《迷失在中间：语言模型如何利用长上下文》，计算语言学协会汇刊，第12卷，第157–173页，2024年。

[18] Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qing-Long Zhang、Xizhou Zhu、Lewei Lu、Bin Li、Ping Luo、Tong Lu、Yu Qiao 和 Jifeng Dai，《InternVL：扩展视觉基础模型并使其适用于通用视觉语言任务》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集 *(CVPR)*，2024年6月。

[19] Zhengxuan Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh，“使用前进行校准：提升语言模型的少样本性能”，载于第 *38* 届国际机器学习大会 *(ICML)*。2021年，第 12697–12706 页，PMLR。

[20] Xuezhi Wang、Jason Wei、Dale Schuur mans、Quoc Le、Ed H. Chi 和 Denny Zhou，“自洽性提升语言模型中的思维链推理”，载于国际学习表征大会 *(ICLR)*，2023年。