Observation Section

1 Observation

理解音频大语言模型（ALLM）如何在内部处理和仲裁相互竞争的情绪线索，是设计有效对抗攻击的前提。本节呈现两组机理发现：§1.1 研究 ALLM 如何解决音频内部韵律与语义情绪信号的冲突；§1.2 研究外部文本指令如何跨模态覆盖音频情绪信号。两组观察共同揭示了一个两级模态优先级层级——音频语义优先于音频韵律（§1.1），文本指令优先于整体音频信号（§1.2）——这一层级结构直接决定了对抗情绪攻击的可行性与设计空间。

1.1 ALLM 如何仲裁韵律-语义冲突

当一段语音的语义内容与韵律传递表达相互矛盾的情绪时（例如，语义悲伤的句子以快乐的语调朗读），ALLM 必须在两种线索之间进行内部仲裁。我们采用三种互补的可解释性工具——线性探针（Probe）、Logit Lens 和激活修补（Activation Patching）——追踪模型 36 层 Transformer（Layer 0–35）中的仲裁过程。核心发现是模型呈现**三阶段层级结构**：早层韵律表征主导、中层融合竞争、晚层语义决策主导。关键在于，韵律信息在早层具有表征可读性，但在晚层不被决策机制采纳。

1.1.1 实验设置

我们构建了包含 247 条音频样本的受控数据集，涵盖五种情绪类别（neutral、happy、sad、angry、surprised）。对 50 条基础文本（每类 10 条），通过 TTS 合成五种韵律变体，得到语义与韵律情绪标签一致（50 条）或冲突（197 条）的样本。Prompt 固定为中性情绪分类指令：

> *"What is the emotion of this audio? Answer with exactly one word: neutral, happy, sad, angry, surprised."*

对每条样本，通过单次前向传播提取所有 36 层的隐状态，对 audio token span 做均值池化得到逐层表示向量。

1.1.2 逐层 Probe 分析

在每一层 ℓ，我们训练两个独立的线性探针：一个预测语义情绪标签（由文本内容决定），一个预测韵律情绪标签（由语音传递方式决定）。定义**主导性指标**：

$$D(\ell) = \mathrm{Acc}_{\mathrm{prosody}}(\ell) - \mathrm{Acc}_{\mathrm{semantic}}(\ell) \tag{1}$$

其中 $D(\ell) > 0$ 表示韵律信息更具线性可读性，$D(\ell) < 0$ 表示语义占优。

Probe 结果揭示了清晰的三阶段结构：

早层（0–14）：韵律主导。 韵律准确率在 Layer 0 达到峰值（$\mathrm{Acc}_{\mathrm{prosody}} \approx 0.842$），在整个区间内持续高于语义准确率。主导性指标均值约为 0.146，在 Layer 5 达到最大值（$D \approx 0.215$）。这表明音频编码器的输出保留了强韵律特征，在早期 Transformer 层中可被轻易解码。

中层（14–23）：融合区。 主导性指标趋近于零（$D \approx 0$），在 Layer 14–15 附近发生符号翻转。在此区域，韵律与语义信息均未明确主导表征，表明两类线索正在进行活跃的竞争与整合。

晚层（23–35）：语义主导。 语义准确率在 Layer 27 达到峰值（$\mathrm{Acc}_{\mathrm{semantic}} \approx 0.830$），主导性指标转为负值（Layer 26 最为显著，$D \approx -0.041$）。在冲突子集（197 条）上，平均韵律准确率为 0.790，语义准确率为 0.730，整体主导性为 0.060——虽然韵律在全局表征层面仍保有一定优势，但晚层向语义主导的转变对模型输出行为具有决定性意义。

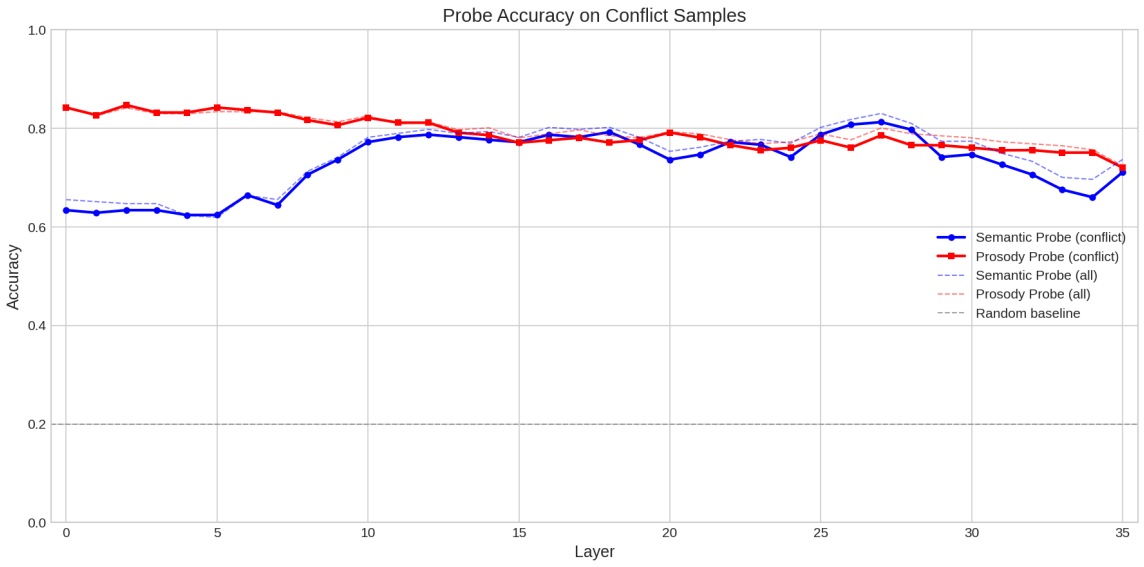

图 1: 逐层 Probe 准确率（冲突样本）。蓝色实线为语义 Probe，红色实线为韵律 Probe。韵律在早层（0–14）主导，语义在晚层（23–35）追平并局部反超。

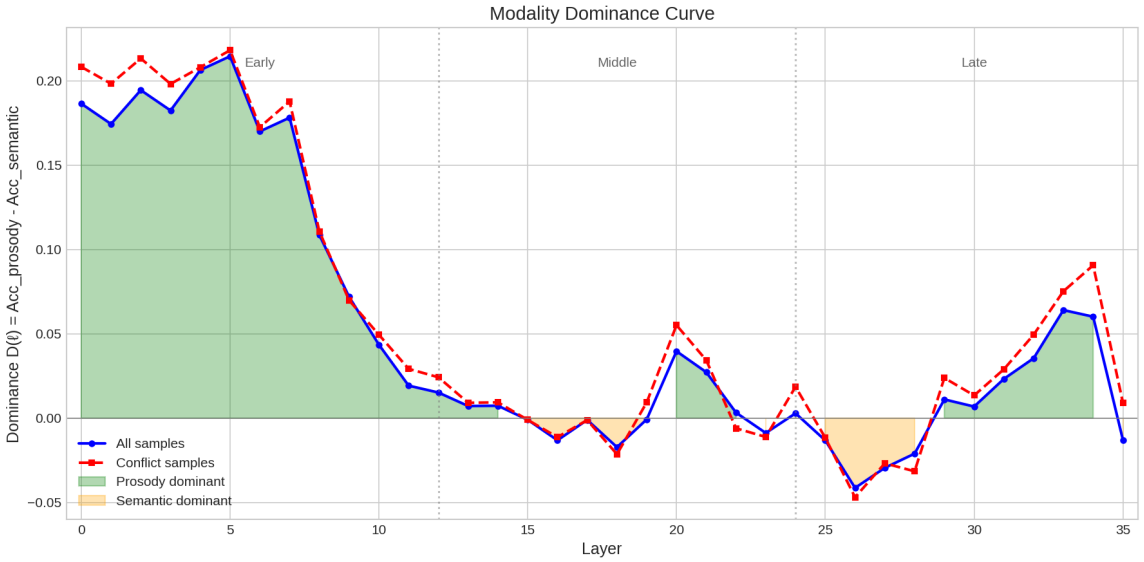

图 2: 模态主导性指标 $D(\ell) = \mathrm{Acc}_{\mathrm{prosody}} - \mathrm{Acc}_{\mathrm{semantic}}$ 逐层变化。绿色填充区域表示韵律主导（$D > 0$），橙色填充区域表示语义主导（$D < 0$）。关键交叉点在 Layer 14–15。

【待补充：Layer 29–34 出现韵律回弹现象，需通过 bootstrap 95% CI 显著性检验 + 非冲突对照组判定是真实机制还是统计伪影。】

【待补充：Probe 稳健性验证——K-fold 交叉验证 + 随机标签对照 + 非冲突负对照，确认探针结果非过拟合。】

1.1.3 Logit Lens：决策轨迹

Probe 揭示了各层编码了什么信息，但并不直接指示什么驱动了模型输出。为弥合这一差距，我们采用 Logit Lens 技术，将每一层在读出位置（最后一个输入 token，$t = T - 1$）的隐状态投影通过最终层归一化和语言模型头，获得中间 logit 分布：

$$\mathrm{logits}_\ell = \mathrm{LMHead}\big(\mathrm{FinalNorm}(\mathbf{h}_\ell^{(T-1)})\big) \tag{2}$$

我们将分析限制在五个情绪 token 上，定义**边际值（margin）**为韵律目标与语义目标 token 的 logit 差：

$$\mathrm{Margin}(\ell) = \mathrm{logit}_\ell(y_{\mathrm{prosody}}) - \mathrm{logit}_\ell(y_{\mathrm{semantic}}) \tag{3}$$

同时计算逐层**胜率（win-rate）**，即冲突样本中语义目标在五个候选中 logit 最高的比例。

Logit Lens 分析揭示了两阶段决策轨迹：

Layer 0–22：决策未成形。 Margin 在零附近波动，大量样本的 argmax 输出为"other"（既非韵律目标也非语义目标）。尽管 Probe 已观察到强韵律可读性，模型在此区间尚未做出决策承诺。

Layer 23–35：语义决策固化。 从 Layer 23 起，margin 持续转负，表明语义目标的 logit 系统性地超过韵律目标。语义胜率在晚层稳步上升。这证明早层韵律表征优势并未转化为决策偏好——模型的输出机制逐步锁定语义解释。

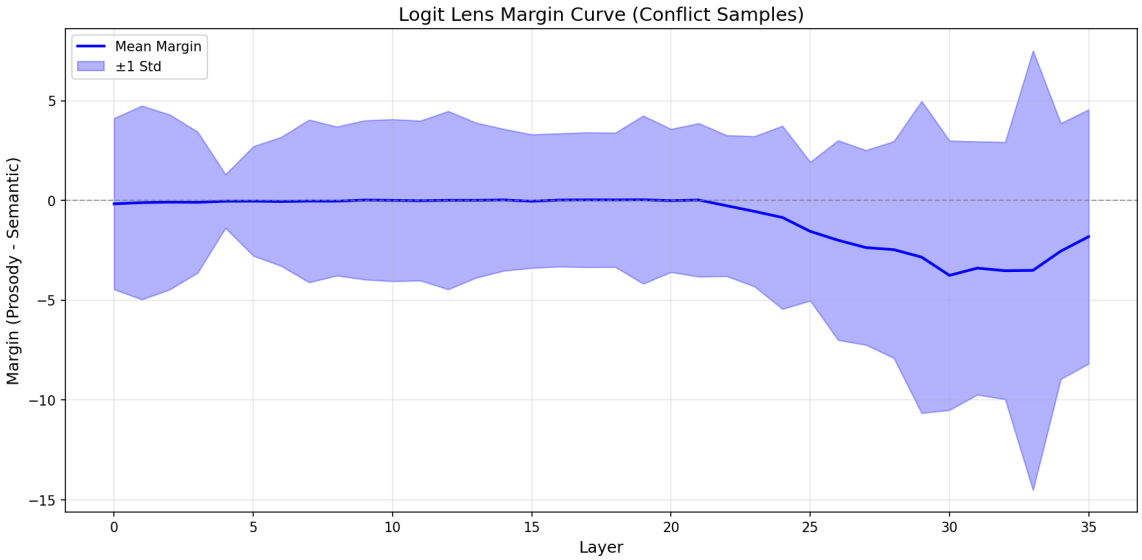

图 3: Logit Lens Margin 曲线（冲突样本）。Margin 定义为韵律目标与语义目标 token 的 logit 差。Layer 0–22 margin 在零附近波动（决策未成形），Layer 23 起持续转负（语义决策固化）。

值得注意的是，决策转折点（Layer 23）明显晚于 Probe 识别的表征交叉点（Layer 14–15），表明决策机制滞后于表征变化。这一"表征先变、决策后变"的时间差（∼8 层）与以下假设一致：中间层的 Logit Lens 投影在表征尚未充分对齐输出嵌入空间时可能不够可靠。

【待补充：词表偏差可能影响 Layer 23 转折的可靠性——需补充 5 词 unigram 基线 + 校正后重绘 margin 曲线，排除词频先验对转折点位置的干扰。】

【待补充：所有指标（$D(\ell)$、margin、win-rate）的 bootstrap 95% CI + 多重比较校正。】

1.1.4 Activation Patching：因果确认

Probe 定位了信息在哪里被编码，Logit Lens 近似了决策何时形成。为确定什么因果性地控制输出，我们采用激活修补（Activation Patching）。对于每对冲突样本 (A, B)（语义或韵律标签不同），在第 ℓ 层将样本 A 的 audio span 隐状态替换为样本 B 的对应表示，然后继续前向传播至输出。我们测量两个指标：

- **Flip-to-Target**：修补后输出匹配 B 的目标标签的样本比例。
- **Delta Logit (Target)**：修补引起的目标 token logit 变化，$\Delta(\ell) = \text{logit}_{\text{patch}}(y_{\text{target}}) - \text{logit}_{\text{base}}(y_{\text{target}})$。

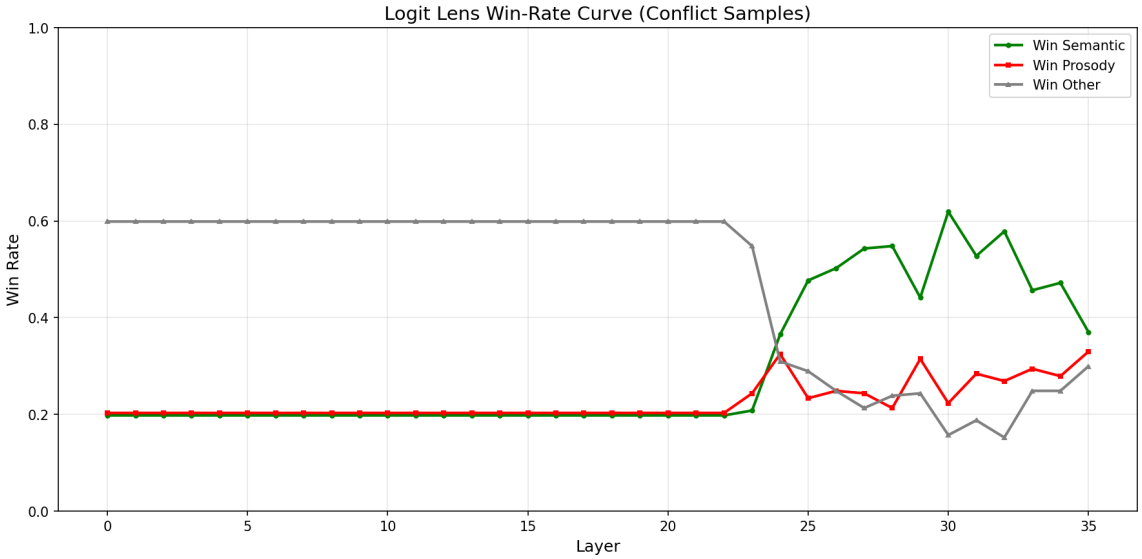

图 4: Logit Lens 逐层胜率（冲突样本）。绿色为语义胜率，红色为韵律胜率，灰色为"other"。Layer 0–22 "other" 占主导（∼0.6），Layer 23 起语义胜率快速上升。

我们分别进行语义修补和韵律修补两类实验：

语义修补（替换语义内容）。 在早期层（0–12），语义修补展现出强因果控制力：Flip-to-Target 最高约 0.65，多层维持在 0.50 以上；Delta Logit 在最早层约为 5–6，随后逐步衰减。这确认了早层 audio 表示中编码的语义内容对最终情绪输出具有强定向因果影响。

韵律修补（替换韵律传递）。 韵律修补的效应显著弱于语义修补：Flip-to-Target 峰值约 0.14–0.26（出现在 Layer 9 附近），Delta Logit 仅约 2.2–2.3。韵律特征的因果影响可测量但远弱于语义特征，与模型最终偏好语义决策的行为一致。

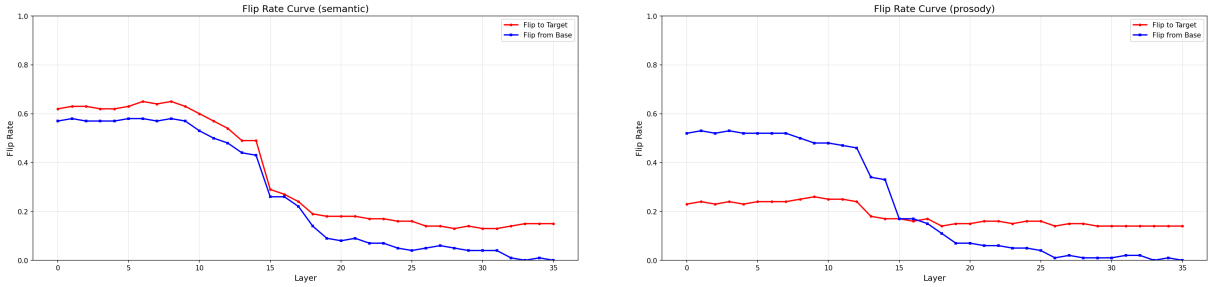

图 5: Activation Patching 逐层 Flip Rate。左：语义修补（Flip-to-Target 在早层 ≈ 0.65），右：韵律修补（Flip-to-Target 峰值仅 ≈ 0.26）。两者均在 Layer 14–15 后急剧衰减。

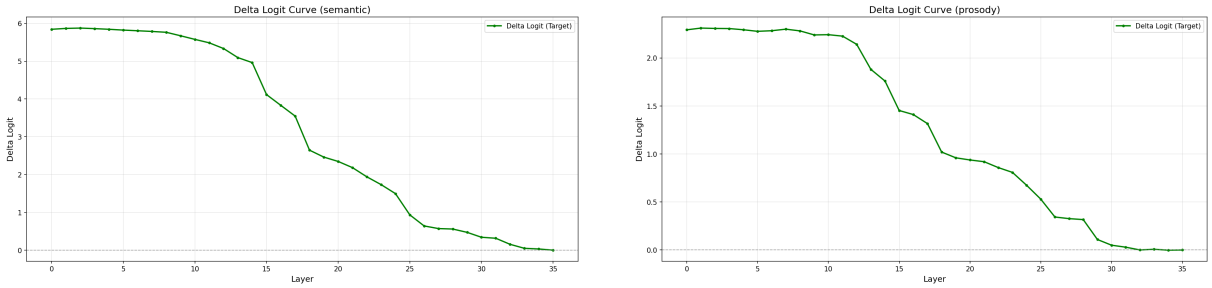

图 6: Activation Patching 逐层 Delta Logit。左：语义修补（早层 $\Delta \approx 5$–6），右：韵律修补（早层 $\Delta \approx 2.2$–2.3）。语义的因果控制力约为韵律的 2.5 倍。

Layer 14–15：可控性边界。 一个关键发现是，超过 Layer 14–15 后，无论是语义还是韵律修补，仅替换 audio token span 对输出的影响均变得微乎其微。这一边界与 Probe 识别的表征交叉点吻合，表明到中层时，情绪相关信息已从局部 audio span 扩散至更广泛的位置表示，使得局部干预失效。

【待补充："信息扩散至全局位置"的解释缺乏直接证据——需补充 position-level patching 实验，对比 audio span / text span / random span 在 Layer 15+ 的控制力变化，验证可控性下降确实源于信息扩散而非其他机制。】

1.1.5 小结

三种工具提供了互补且收敛的证据，共同确认 ALLM 内部的**语义优先仲裁**策略。Probe 揭示*编码了什么*：韵律主导早层表征，语义在晚层接管。Logit Lens 揭示*什么驱动输出*：决策轨迹从 Layer 23 起锁定语义解释。Activation Patching 揭示*什么因果性地决定输出*：语义内容具有强定向控制力（Flip-to-Target ≈ 0.65），韵律仅具弱扰动效应（Flip-to-Target ≈ 0.14–0.26）。

两个关键边界浮现：*表征交叉*在 Layer 14–15（编码从韵律主导转向语义主导），*决策转折*在 Layer 23（输出机制锁定语义）。两者之间约 8 层的间隔反映了表征变化与决策承诺之间的滞后。

【待补充：三组对照对齐——Audio-only / Conflict / Consistent 的 Probe、Logit Lens、Activation Patching 结果需并列呈现，避免"单组现象"被误读为普遍机制。】

【待补充：Layer 14–15 vs Layer 23 边界落差的解释——是"表征先变、决策滞后"的真实机制，还是 Logit Lens 在中间层投影不可靠的伪影？需设计实验区分。】

1.2 文本指令如何覆盖音频情绪信号

§1.1 考察了音频内部的模态冲突。本节转向跨模态冲突：当文本指令（prompt）将模型引向与音频信号矛盾的情绪时，会发生什么？这一场景直接关系到对抗鲁棒性，因为它揭示了 ALLM 决策过程中文本与音频模态之间的结构性优先级。核心发现是：文本指令在**中层（Layer 5–20）建立强因果主导**，结构性地压制音频情绪信号，最终决

策在**Layer 26–28 不可逆地固化**为文本指向。

1.2.1 实验设计

为隔离文本指令情绪与音频情绪之间的纯冲突——排除语义内容变化的干扰——我们采用以下受控设计。

数据构建。 使用 TTS 生成固定语义内容的音频（如 "The package is scheduled to arrive..."），以五种不同情绪韵律朗读。语义文本内容恒定，确保观察到的任何冲突纯粹来自 prompt 中的指令情绪指向与音频中的韵律情绪之间的对立，排除语义内容作为混淆变量。实验涵盖 18 条基础音频 $\times$ 3 种条件。

三组实验条件。

- **Audio-only**：prompt 仅要求判断音频情绪，不指定目标方向。
- **Conflict**：prompt 指令模型以特定目标情绪 T 判断，其中 T 与音频实际韵律情绪 A 不同（$T \neq A$）。
- **Consistent**：prompt 指令指向 $T = A$，作为正对照。

分析方法。 沿用 §1.1 的两种工具：

1. **Logit Lens 差分**：计算逐层 logit 差值 $\Delta\text{logit}_\ell = \text{logit}_\ell(T) - \text{logit}_\ell(A)$，定位仲裁发生在哪里。
2. **Activation Patching**：分别修补文本 token 和音频 token 的隐状态，确定哪个模态因果性地驱动输出。

1.2.2 发现一：晚期层决策固化

对三组条件的 Logit Lens 差分分析揭示，文本指令与音频情绪之间的仲裁是一个晚期层现象，且存在不可逆的固化点。

Layer 0–20：无分化。 在所有三组条件下，逐层 logit 差分 Δlogit_ℓ 在前 20 层均接近零。无论文本指令是否与音频情绪一致，模型在此区间尚未开始解决跨模态冲突。

Layer 26–28：决策结晶。 在 Layer 26–28 出现急剧分化。Conflict 条件下，Δlogit_ℓ 决定性地偏向文本指令情绪 T；Consistent 条件下，两种模态相互增强同一目标。关键在于，一旦决策在此窄窗口内结晶，后续层不再出现逆转——决策在 Layer 28 之后**不可逆**。这定义了一个有限的决策窗口：模型的情绪输出在约 2–3 层的跨度内被确定，此后不再发生模态竞争。

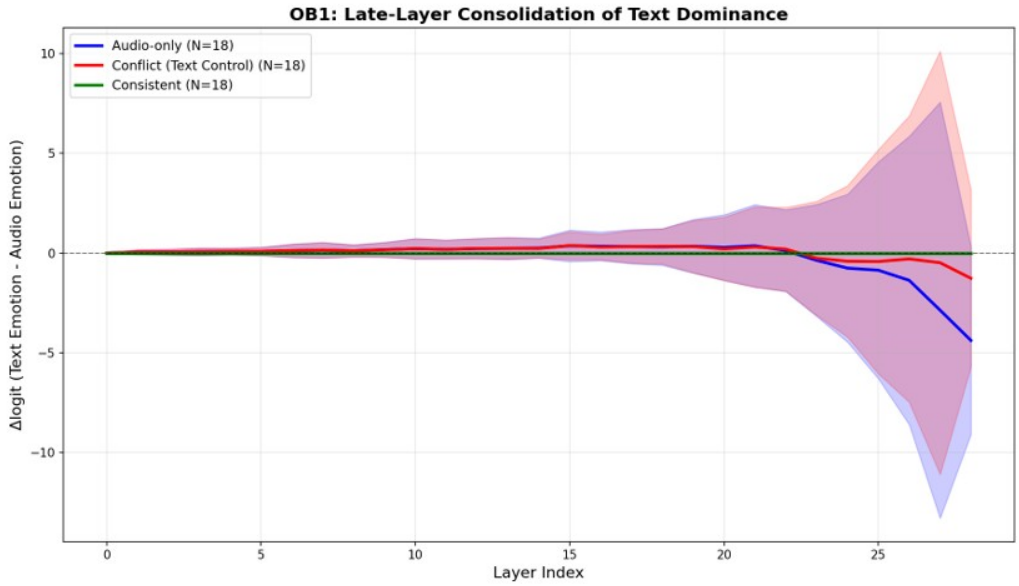

图 7: 跨模态 Logit Lens 差分（$\Delta\text{logit}_\ell = \text{logit}_\ell(T) - \text{logit}_\ell(A)$）。三组条件（Audio-only / Conflict / Consistent）在前 20 层无分化，Layer 26–28 出现决策结晶。阴影区域为 ±1 标准差。

【待补充：仲裁固化的不可逆性验证——需在 Layer 26–28 之后做 PatchAudio，确认对输出无影响，从而严格证明决策窗口的有限性。】

1.2.3 发现二：文本因果主导，音频中层被结构性压制

Logit Lens 定位了仲裁在哪里发生，Activation Patching 进一步揭示了谁在因果层面主导这一仲裁。我们分别在各层修补文本 token（PatchText）和音频 token（PatchAudio）的隐状态，比较两者对最终输出的因果效应。

PatchText 在 Layer 5–20 有效。 将文本指令的隐状态替换为指向不同情绪的版本后，模型的最终情绪输出发生显著改变。文本指令在 Layer 5–20 的广泛区间内保持强因果效应，表明文本模态在中层即已建立对决策的主导控制。

PatchAudio 在相同层段无效或不稳定。 在同一层段（Layer 5–20）替换音频隐状态，对最终输出的影响微弱且不稳定。音频情绪信号在中层被结构性压制（structural wash-out）——并非简单的加权竞争，而是文本通路在中层系统性地"覆盖"了音频信号的因果贡献。

因果不对称性。 PatchText 与 PatchAudio 的效应对比揭示了显著的**因果不对称性**：在决策形成的关键层段，文本指令具有强定向因果控制力，而音频信号的因果贡献被边缘化。这不是两种模态的对等竞争，而是文本模态的结构性主导。

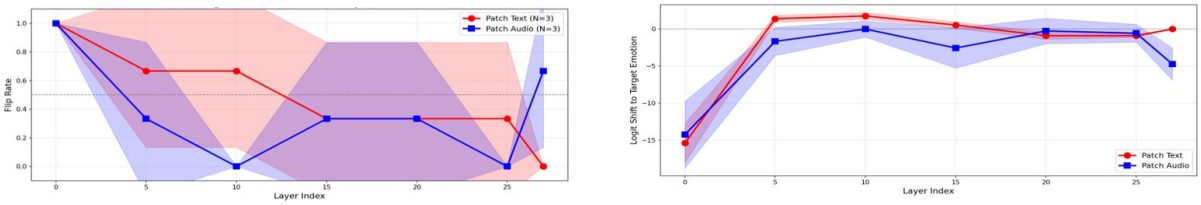

图 8: 跨模态 Activation Patching 对比。左：Flip Rate（PatchText 红色 vs PatchAudio 蓝色）。右：Logit Shift to Target Emotion。PatchText 在 Layer 5–20 持续有效，PatchAudio 在相同区间无效或不稳定，揭示文本的因果不对称主导。

【待补充：文本主导性的量化指标（Text-Dominance Index）——在 Layer 5–20 对 text token 做 mean-ablation 测量输出变化，定义为标量指标，与音频 ablation 效应做定量对比。】

【待补充：音频信号在中层被"压制"的具体机制——是 attention 权重重分配？还是 MLP 层的非线性消解？需通过 attention pattern 分析或 MLP 贡献分解进一步确认。】

【待补充：不同指令复杂度对中层主导性的影响——简单指令 vs 多步指令（如 CoT 格式）的 PatchText 因果强度对比。】

1.2.4 与 §1.1 的关系：两级模态优先级

将 §1.1 和 §1.2 的发现联合分析，揭示了 ALLM 情绪决策中的**两级模态优先级层级**：

第一级：音频内语义 > 音频内韵律（§1.1）。 当语义与韵律冲突时，模型在 Layer 23 起将决策锁定为语义解释。语义修补的因果控制力（Flip-to-Target ≈ 0.65）远超韵律修补（≈ 0.14–0.26）。

第二级：文本指令 > 音频整体信号（§1.2）。 当外部文本指令参与竞争时，文本的因果主导性更强、作用区间更广（Layer 5–20），相比 §1.1 中语义的决策转折点（Layer 23）更早介入。文本指令不仅覆盖音频韵律，还覆盖音频语义——实现了对整体音频信号的结构性压制。

层级边界的嵌套关系。 §1.1 中识别的 audio span 可控性边界（Layer 14–15）恰好落在 §1.2 中文本主导区间（Layer 5–20）内。这一嵌套关系暗示：一旦文本在中层建立控制，音频侧的任何扰动——无论针对语义还是韵律——都难以穿越此区间影响最终决策。

1.2.5 过渡：从机理发现到攻击设计

§1.1–§1.2 的机理发现表明，ALLM 的情绪决策存在清晰的层级结构和模态优先级。具体而言：

1. 音频内部的情绪仲裁遵循"语义优先"策略，韵律信息虽在早层可读但不被决策采纳；
2. 文本指令在中层即建立因果主导，结构性地压制音频信号，决策在 Layer 26–28 不可逆固化；
3. 两级优先级（音频语义 > 韵律；文本 > 音频）共同约束了音频对抗扰动的作用空间。

这些发现为理解音频对抗扰动的可行性边界与限制条件提供了机理基础，自然引出以下问题：在上述层级约束下，对抗扰动能否以及如何突破模型的仲裁机制实现定向情绪攻击？这一问题将在 Section 3 中通过攻击方法论设计与实验验证予以回答。

【待补充：跨模型验证——Prompt-Audio 冲突实验在 Kimi-Audio 等其他 ALLM 上的复现，验证上述机理发现的普适性。】