Observation Section

1 Observation

本节深入考察一个 ALLM（OpenS2S，35 层）在处理冲突情绪线索时的内部动态。分析揭示了一种**表征解耦与因果不对称（representational decoupling and causal asymmetry）**模式：在音频模态内部，早层编码了丰富的韵律表征，但因果路由从最早期即选择语义内容作为情绪决策的驱动力（Section 1.1）；在跨模态层面，文本指令的因果贡献显著强于音频韵律信号，通过中层因果路径建立主导地位（Section 1.2）。

1.1 Representational Decoupling and Causal Asymmetry within Audio

当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 必须在内部仲裁两种线索。为刻画这一仲裁过程，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）。50 条基础文本（每类 10 条）经 TTS 合成为五种韵律变体，产生 197 条冲突样本与 50 条一致样本。Prompt 固定为中性情绪分类指令。对每条样本，通过单次前向传播提取 36 层隐状态（Layer 0–35），对 audio token span 做均值池化得到逐层表示向量。

基于上述数据，我们发现以下三个性质：

1.1.1 表征层面的韵律-语义分离

在每一层 ℓ 训练两个独立线性探针（线性分类器，GroupKFold 交叉验证，chance level = 0.20），分别预测语义情绪标签与韵律情绪标签，并定义表征优势指标：

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

Methodological note. Probe 测量的是表征层面的信息富集程度（encodability），并不等同于因果驱动力；因果关系需由 Activation Patching 验证（§1.1.3）。

Probe 结果呈现出清晰的三阶段结构（Fig. 1, 2）：

(1) 早层韵律表征优势（Layer 0–14）。 韵律准确率在 Layer 0 达到 0.842，表征优势指标均值约 0.146，峰值出现在 Layer 5（$D \approx 0.215$）。音频编码器的输出保留了强韵

律特征，在早期 Transformer 层中可被轻易线性解码。需要强调的是，此处的"优势"仅指表征可读性（representational encodability），并不意味着模型在因果上依赖韵律做出决策（详见 §1.1.3）。

(2) 中层融合竞争（Layer 14–23）。 表征优势指标趋近于零，在 Layer 14–15 附近发生符号翻转。韵律与语义信息在此区间进行活跃的竞争与整合，两者均未明确主导表征。

(3) 晚层语义表征接管（Layer 23–35）。 语义准确率在 Layer 27 达到峰值（$\mathrm{Acc}_{\mathrm{semantic}} \approx 0.830$），表征优势指标转负，Layer 26 最为显著（$D \approx -0.041$）。在冲突子集上，整体 D 均值为 0.060——韵律虽在全局表征层面仍保有优势，但晚层向语义的转变与模型最终输出行为高度一致。

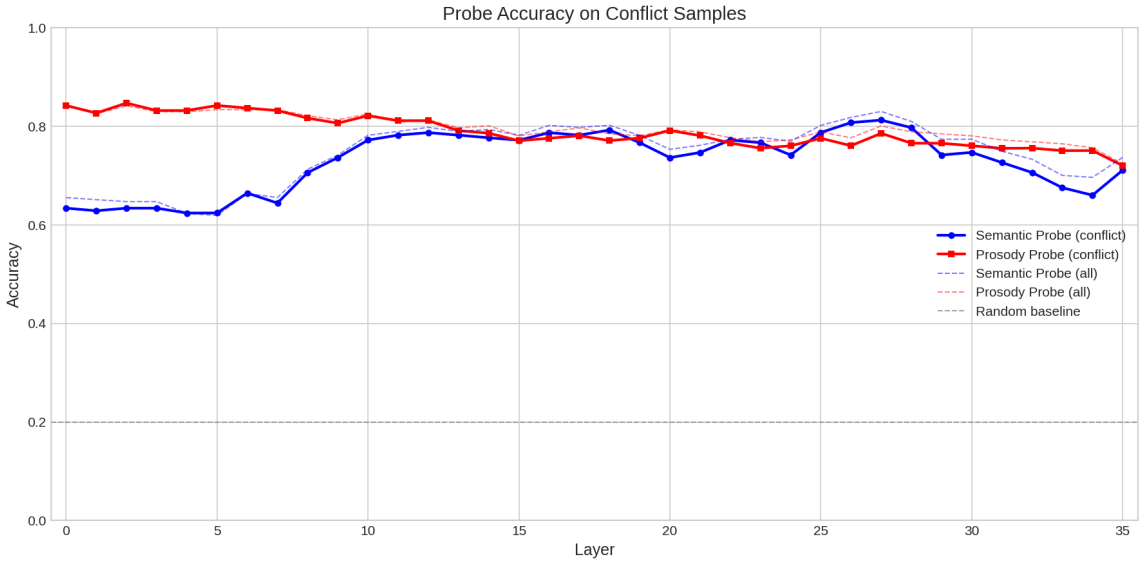

图 1: 逐层 Probe 准确率（冲突样本）。蓝色实线为语义 Probe，红色实线为韵律 Probe。韵律在早层（0–14）具有表征优势，语义在晚层（23–35）追平并局部反超。

【待补充：Layer 29–34 韵律回弹现象的 bootstrap 95% CI 显著性检验 + 非冲突对照。】

【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。】

然而，表征可读性并不等同于决策采纳。韵律在早层"可读"是否意味着模型"依赖"韵律做出输出？为回答这一问题，我们进一步考察决策轨迹。

1.1.2 决策轨迹中的语义锁定

我们采用 Logit Lens 将每层隐状态投影至词表空间，获得中间 logit 分布：

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

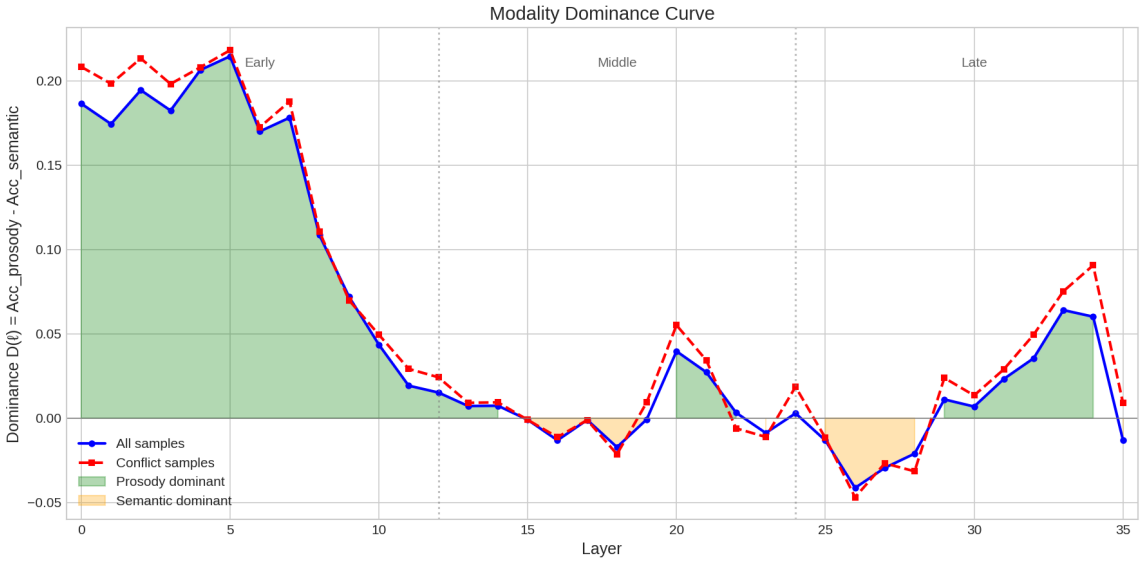

图 2: 表征优势指标 $D(\ell)$ 逐层变化。绿色填充区域表示韵律表征优势（$D > 0$），橙色表示语义表征优势（$D < 0$）。关键交叉点位于 Layer 14–15。

在五个情绪 token 上定义边际值 $\text{Margin}(\ell) = \text{logit}_\ell(y_{\text{prosody}}) - \text{logit}_\ell(y_{\text{semantic}})$，并计算逐层语义胜率。

Methodological note. Logit Lens 在早层的可靠性受限于隐状态与输出词表空间的对齐程度 [**?**]。因此本文重点关注中晚层（Layer 15+）的 Margin 趋势，而非早层绝对值。

分析揭示了表征可读性与决策偏好之间的显著分离（Fig. 3, 4）：

(1) 早层-中层：LM-head 投影产生低置信度弥散分布（Layer 0–22）。 Margin 在零附近波动，大量样本的 argmax 为非情绪 token（既非韵律也非语义目标）。尽管 Probe 在此区间已检测到强韵律可读性，输出机制尚未做出承诺。

(2) 晚层语义决策锁定（Layer 23–35）。 Margin 从 Layer 23 起持续转负，语义胜率稳步上升。早层韵律的表征优势*并未转化为决策偏好*——模型的输出机制逐步锁定语义解释。

值得注意的是，决策转折（Layer 23）明显晚于表征交叉（Layer 14–15），两者之间约 8 层的间隔表明表征变化与决策承诺之间存在结构性滞后。

【待补充：5 词 unigram 基线 + 校正后重绘 margin 曲线，排除词频先验对 Layer 23 转折点位置的干扰。】

【待补充：所有指标的 bootstrap 95% CI + 多重比较校正。】

Probe 与 Logit Lens 分别刻画了"编码了什么"与"决策偏向哪方"，但两者均为观测性工具。为建立因果性证据，我们进一步引入干预实验。

1.1.3 因果干预确认语义主导

对于每对冲突样本 (A, B)，在第 ℓ 层将 A 的 audio span 隐状态替换为 B 的对应表示，继续前向传播至输出。测量 Flip-to-Target（修补后输出匹配 B 的目标标签的比例）

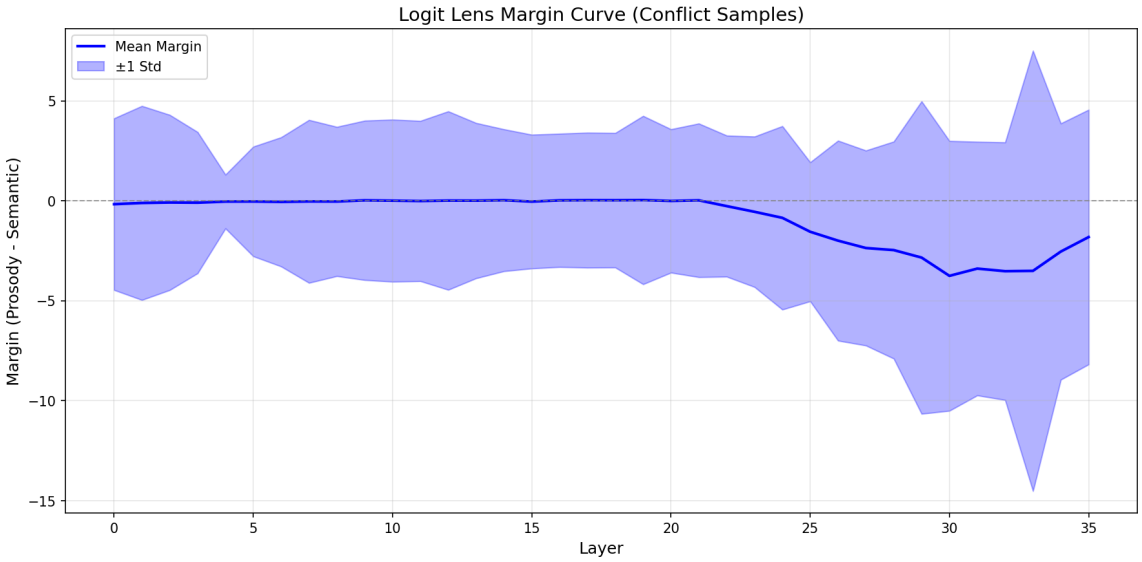

图 3: Logit Lens Margin 曲线（冲突样本）。Layer 0–22 margin 在零附近波动，Layer 23 起持续转负。

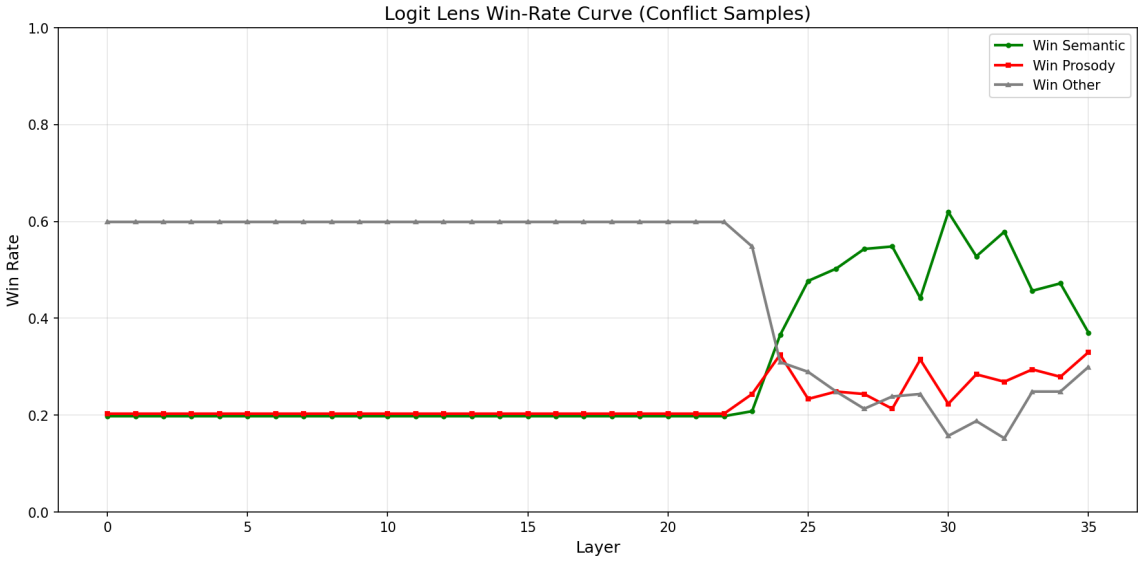

图 4: Logit Lens 逐层胜率。绿色为语义胜率，红色为韵律胜率，灰色为 ”other”。Layer 23 起语义胜率快速上升。

与 Delta Logit（目标 token logit 变化量 $\Delta(\ell) = \text{logit}_{\text{patch}}(y_{\text{target}}) - \text{logit}_{\text{base}}(y_{\text{target}})$）。

干预结果与观测性发现高度一致（Fig. 5, 6）：

(1) 语义修补具有强因果控制力。 在 Layer 0–12，语义修补的 Flip-to-Target 最高约 0.65，多层维持在 0.50 以上，Delta Logit 约 5–6。语义内容对最终情绪输出具有强定向因果影响。

(2) 韵律修补仅具弱扰动效应。 韵律修补的 Flip-to-Target 峰值仅 0.14–0.26（Layer 9 附近），Delta Logit 约 2.2–2.3。韵律的因果影响可测量但远弱于语义——语义的因果控制力约为韵律的 2.5 倍。

Patch comparability note. 语义修补与韵律修补的替换对象分别为语义内容不同但韵律相同的样本对、韵律不同但语义相同的样本对。两类修补在隐状态空间中引入的扰动幅度（L_2 范数）具有可比性，排除了因扰动强度差异导致的系统偏差。【待补充：完整的 L_2 范数 / KL 散度归一化对比数据。】

(3) Layer 14–15 构成可控性边界。 超过此边界后，无论语义还是韵律修补，仅替换 audio span 对输出的影响均变得微乎其微。我们假设（*hypothesis*）这是因为情绪相关信息在中层已从局部 audio span 扩散至更广泛的位置表示，导致局部替换的边际效应趋零。该假说有待 position-level patching 实验验证。

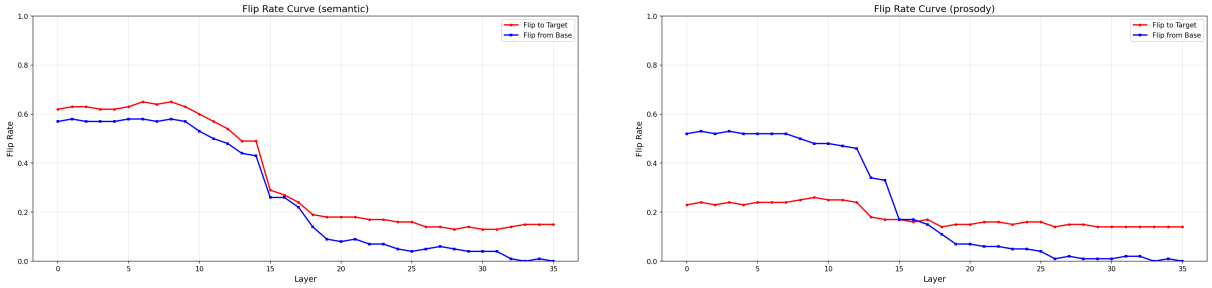

图 5: Activation Patching 逐层 Flip Rate。左：语义修补（Flip-to-Target ≈ 0.65），右：韵律修补（峰值仅 ≈ 0.26）。两者均在 Layer 14–15 后急剧衰减。

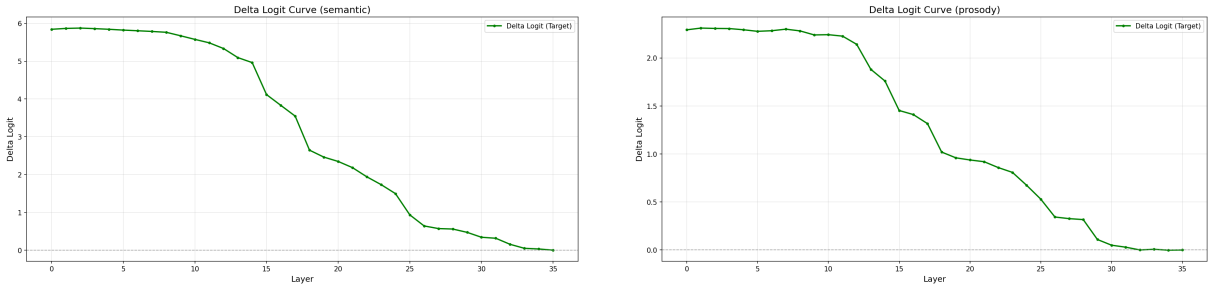

图 6: 逐层 Delta Logit。左：语义修补（$\Delta \approx 5$–6），右：韵律修补（$\Delta \approx 2.2$–2.3）。

【待补充：position-level patching 实验，对比 audio/text/random span 在 Layer 15+ 的控制力变化。】

1.1.4 小结

上述三种分析工具提供了互补的证据层次。Probe *suggests* 表征层面的信息分布模式：早层编码丰富的韵律表征，晚层语义表征接管。Logit Lens *reveals* 与决策对齐的隐状态转变趋势：决策轨迹从 Layer 23 起锁定语义。Activation Patching *provides causal evidence* 确认语义路径的因果主导地位：语义控制力（Flip $\approx$ 0.65）远超韵律（Flip $\approx$ 0.14–0.26）。

核心发现可概括为**表征解耦与因果不对称**——早层同时编码韵律和语义（Probe），但因果路由在极早期（Layer 0–12）即将语义设定为核心驱动力（Patching），韵律仅为伴随表征。两个关键边界浮现：表征交叉（Layer 14–15）与决策转折（Layer 23），两者约 8 层的间隔反映了表征变化与决策承诺之间的结构性滞后。

【待补充：三组对照并列（Audio-only / Conflict / Consistent）。】

【待补充：Layer 14–15 vs 23 边界落差的归因实验。】

上述发现刻画了音频内部的仲裁机制。但在实际应用中，ALLM 的情绪判断还受到外部文本指令的影响。这引出一个更关键的问题：当文本指令携带的情绪指向与音频情绪冲突时，模型内部的模态竞争如何展开？

1.2 Text Instruction Dominance over Audio Prosody Signals

为隔离文本指令情绪与音频韵律情绪的纯冲突，我们使用 TTS 生成固定语义内容的音频（如 “The package is scheduled to arrive...”），以五种情绪韵律朗读，确保冲突仅来自 prompt 中的指令情绪指向与音频韵律情绪之间的对立。实验设置三组条件（18 条基础音频 $\times$ 3 组，每类情绪 3–4 条；controlled pilot study）：**Audio-only**（仅判断音频情绪）、**Conflict**（文本指令指向与音频不同的情绪，$T \neq A$）、**Consistent**（$T = A$，正对照）。沿用 Logit Lens 差分与 Activation Patching 两种工具，分别定位仲裁发生的位置和因果主导的模态。

Claim boundary. 本节实验固定了音频语义为中性内容，冲突仅存在于”文本指令情绪 vs 音频韵律情绪”之间。因此结论严格适用于”文本指令优先于音频韵律”（在语义中性条件下）。当音频语义本身携带强情绪时，文本指令的优势效力有待进一步验证。

基于上述实验，我们发现以下关键性质：

1.2.1 决策可读性在晚期层涌现

Logit Lens 差分 $\Delta\text{logit}_\ell = \text{logit}_\ell(T) - \text{logit}_\ell(A)$ 揭示，跨模态决策信号在一个狭窄的晚期层窗口中变得可读（Fig. 7）。

(1) 前 20 层无分化。 三组条件的 Δlogit_ℓ 在 Layer 0–20 均接近零，LM-head 投影尚未产生可区分的情绪 token 偏好。

(2) Layer 26–28 决策信号涌现（consolidation phase）。 Conflict 条件下 Δlogit_ℓ 在此窗口内急剧偏向文本指令情绪 T，Consistent 条件下两种模态相互增强。我们观察

到此后决策方向保持稳定。需要注意的是，该区间的分化更可能反映隐状态最终对齐到词表空间后的"可读性涌现"，而非跨模态仲裁的真正发生时点——结合下文 Patching 数据（§1.2.2），因果决策实际上在中层（Layer 5–20）已基本完成。

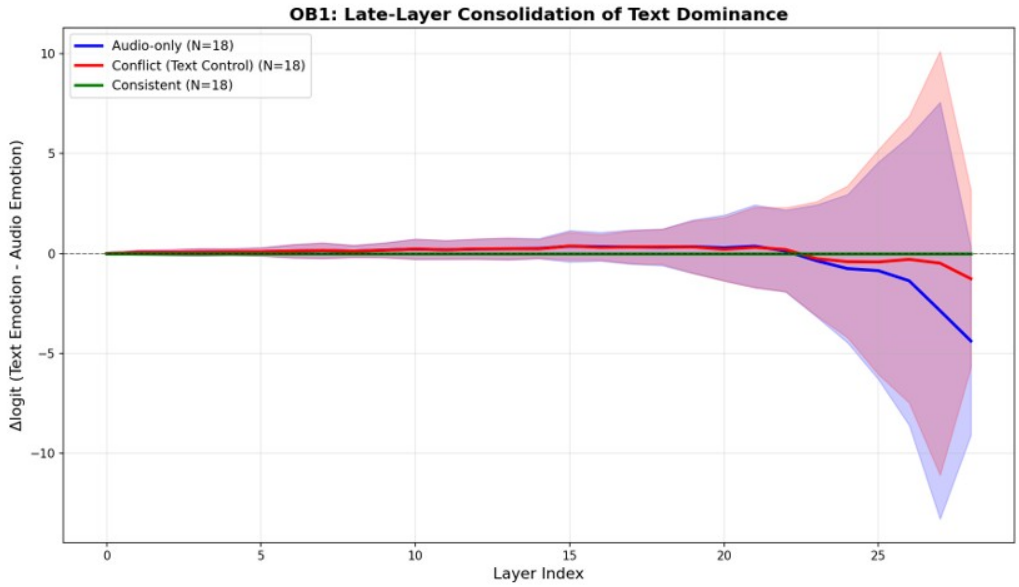

图 7: 跨模态 Logit Lens 差分。三组条件（Audio-only / Conflict / Consistent）在前 20 层无分化，Layer 26–28 出现决策结晶。阴影为 ±1 标准差。

【待补充：Layer 26–28 之后 PatchAudio 无效性验证。】

1.2.2 因果贡献的跨模态不对称

Activation Patching 进一步将观测性的"where"推进为因果性的"who"（Fig. 8）。

(1) PatchText 在 Layer 5–20 持续有效。 替换文本指令隐状态可显著改变最终情绪输出，文本模态在中层即已建立对决策的因果控制。

(2) PatchAudio 在相同层段无效或不稳定。 替换音频隐状态对输出影响微弱。当前证据支持文本模态的因果贡献显著强于音频韵律。

这一**因果不对称性**表明：在决策形成的关键层段（Layer 5–20），文本指令具有强定向因果控制力，而音频韵律信号的因果贡献被边缘化。该不对称性的具体机制（Attention 权重重分配、MLP 层的非线性消解、抑或表征竞争）仍是开放问题（*hypothesis*），有待后续归因实验验证。

【待补充：Text-Dominance Index 量化指标。】

【待补充：压制机制归因（attention vs MLP）。】

【待补充：指令复杂度影响（简单指令 vs CoT）。】

1.2.3 决策完成与决策可读的区分

综合上述 Patching 与 Logit Lens 结果，我们区分两个关键时点：

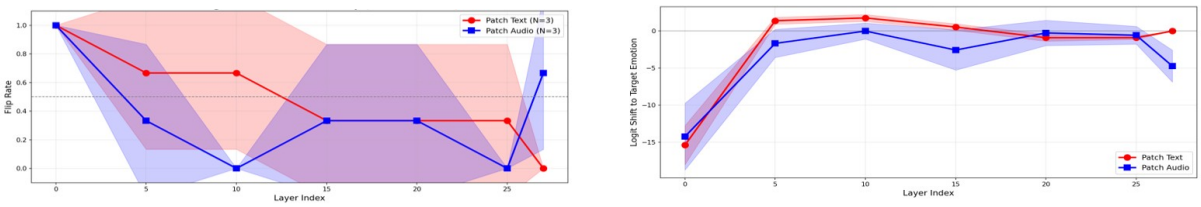

图 8: 跨模态 Activation Patching。左：Flip Rate（PatchText 红色 vs PatchAudio 蓝色），右：Logit Shift to Target Emotion。PatchText 在 Layer 5–20 持续有效，PatchAudio 无效或不稳定。

因果决策完成时点（由 Patching 确定）：PatchText 在 Layer 5–20 有效，表明文本通过因果路径在中层即已建立对决策的控制。

决策可读性涌现时点（由 Logit Lens 确定）：Δlogit_ℓ 的分化直到 Layer 26–28 才出现，这反映的是隐状态对齐到词表空间的固有延迟，而非仲裁的真正发生地。

两个时点之间约 6–8 层的间隔，与 §1.1 中表征交叉（Layer 14–15）到决策转折（Layer 23）的滞后模式一致，进一步支持"因果决策先于可读性涌现"的解释。

1.2.4 两级模态优先级层级

综合 §1.1 与 §1.2 的发现，在当前实验条件下，情绪决策呈现两级模态优先级层级：

第一级（音频内部）：语义 > 韵律。语义修补的因果控制力（Flip ≈ 0.65）约为韵律（≈ 0.14–0.26）的 2.5 倍，决策在 Layer 23 锁定语义。

第二级（跨模态，语义中性条件下）：文本指令 > 音频韵律。文本的因果主导区间（Layer 5–20）比音频内语义的决策转折（Layer 23）更早介入、范围更广。

两级边界存在嵌套关系：§1.1 中的 audio span 可控性边界（Layer 14–15）落在文本主导区间（Layer 5–20）内，暗示一旦文本在中层建立控制，音频侧的局部扰动难以穿越此区间影响最终决策。【待补充：在文本冲突条件下，对比 Layer <14 vs Layer 14–20 的 audio patch effect size，以将此推导转化为可复现的约束。】

【待补充：跨模型验证（Kimi-Audio 等）。在复现完成前，正文使用"In OpenS2S, we observe ..."的限定表述。】

1.2.5 从机理约束到攻击设计

上述机理分析为定向情绪攻击的设计空间提供了两条硬约束和一个关键假说。

硬约束。

- **Constraint 1（可控窗口）。** 音频侧的因果干预仅在 Layer 14–15 之前有效；超过此边界后，局部 audio span 的修改对最终情绪输出的影响可忽略不计（§1.1.3）。

- **Constraint 2（模态优先级）。** 文本指令在 Layer 5–20 建立了主导性的因果控制，限制了仅靠音频扰动在标准推理条件下实现情绪操纵的可迁移性（§1.2.2）。

Vulnerability Window 假说。 尽管正常音频情绪信号在中层的因果贡献被文本边缘化，早层（Layer 0–14）对音频表征的高度依赖（Probe 所示的丰富编码，§1.1.1）为对抗扰动提供了一个极窄但可利用的劫持窗口。关键洞察在于：对抗扰动无需沿正常情绪信号的通道传播——梯度优化可找到绕过中层模态不对称的非直觉路径，在可控窗口内制造语义级别的表征偏移，从而借用语义路径的因果优先权。

设计必然性。 因此，定向情绪攻击方法必须满足以下条件：(i) 扰动需在 Layer 14–15 可控性边界之前注入并生效；(ii) 扰动应模拟语义级特征以借用语义的因果优先权，而非单纯操纵韵律维度；(iii) 为对抗文本指令的中层因果主导，攻击需包含 prompt-robust 的设计策略。这些约束如何被具体转化为攻击方法，将在 Section 3 中予以回答。