LAYERREFIT: FINAL-RELATIVE REFINEMENT FOR AFFECTIVE MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Encoder hierarchies contain affective cues whose utility varies across modalities and conditions. We introduce **LayerRefit** to adapt pretrained multimodal large language models (MLLMs) by converting this evidence into residual corrections of their native modality tokens. Depth-specific evidence projections transform differences from the final encoder state; a shared correction projection maps them into token updates, combined with input-dependent weights. Backbone weights and input sequence length remain fixed. Across three affective backbones, source-trained branches improve 22 of 27 cross-benchmark evaluations. Models trained with audio, vision, and text improve 11 of 12 inference-time input configurations on the two primary backbones without retraining, using fewer than 0.6M added parameters. Ablations favor final-relative evidence with shared token correction, and norm-matched interventions show that learned update directions improve recognition. Code: `https://anonymous.4open.science/r/LayerRefit-7263/`.

1 INTRODUCTION

Affective multimodal large language models (MLLMs) learn to interpret acoustic and visual evidence through multimodal alignment and task-oriented supervision (Cheng et al., 2024; Lian et al., 2025; Peng et al., 2026). This training establishes an interface through which modality tokens support emotion recognition. We study how information distributed across the encoder hierarchy can improve these tokens while reusing the learned backbone.

Encoder representations capture different acoustic, visual, and paralinguistic properties across depth (Pasad et al., 2021; 2023; Saliba et al., 2024; Chen et al., 2025; Pang et al., 2026). The relevance of prosody and facial expression also varies with affective conditions (Juslin & Laukka, 2003; Scherer & Ellgring, 2007). Our HuBERT audio (Hsu et al., 2021) and ViT visual (Dosovitskiy et al., 2021) probes across six affective datasets (Lian et al., 2023; 2024; Poria et al., 2019; Busso et al., 2008; Bagher Zadeh et al., 2018; Yu et al., 2020) reveal modality- and benchmark-dependent preferred depths (Figure 1(a,b)). Within CREMA-D (Cao et al., 2014), the relative utility of intermediate depths also varies across emotions and expression intensities (Figure 1(c)). These patterns motivate using encoder hierarchy as conditional evidence for adaptation, with contributions that vary across depths and inputs.

Figure 2(a,b) illustrates two routes for incorporating encoder hierarchy into the LLM input. *Depth Tokenization* introduces additional tokens derived from intermediate encoder states, as in Dense Connector's sparse token integration (Yao et al., 2024) and LION's visual aggregation branch (Chen et al., 2024a). *Depth Fusion* integrates multi-layer evidence into a fixed-length token sequence through cross-layer attention (Cao et al., 2024) or instruction-conditioned weighting (Li et al., 2025; Lin et al., 2026). For a pretrained affective model, we study residual correction of the native modality tokens produced by its existing connector. The central design problem is to translate heterogeneous depth evidence into token corrections that improve emotion discrimination through the existing interface.

We introduce **LayerRefit**, which predicts residual updates to native modality tokens from encoder-depth evidence (Figure 2(c)). Each modality uses the final encoder state as a common reference. *Depth-specific evidence projections* transform intermediate states relative to this reference, and a *shared correction projection* maps the resulting evidence into token updates. Input-dependent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

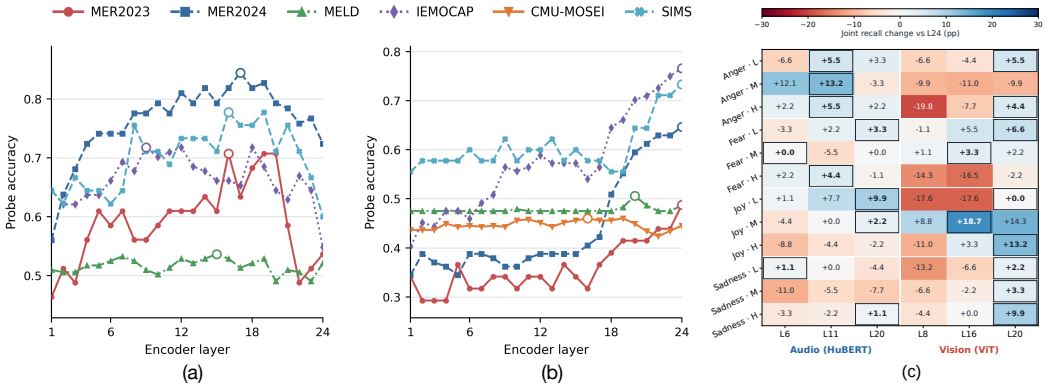


Figure 1: **Depth-dependent affective readouts across benchmarks and conditions.** (a,b) Across benchmarks: probing accuracy over HuBERT audio and ViT visual layers, respectively; hollow circles mark each benchmark's peak probe layer. (c) Within CREMA-D: joint recall changes relative to the final layer (L24), grouped by emotion and expression intensity (L/M/H: low/medium/high). Blue/red indicate gains/losses in percentage points; outlines mark a row maximum among the displayed layers within each modality.

weights combine these updates. Layer-specific transformations accommodate differences in the evidence, while shared output directions coordinate its effect on the tokens within a low-dimensional subspace. The encoders, connector, and LLM remain frozen, with unchanged token counts.

We evaluate this design on AffectGPT (Lian et al., 2025), Emotion-LLaMAv2 (Peng et al., 2026), and Light-MER (Zheng et al., 2026). Source-trained branches improve 22 of 27 cross-benchmark evaluations. On the two primary backbones, models trained with audio, vision, and text improve 11 of 12 inference-time input configurations without retraining, using fewer than 0.6M added parameters. Comparisons with final-layer-only refinement test the value of depth evidence; ablations and directional interventions examine how that evidence becomes an effective correction.

Our contributions are threefold:

1. **Hierarchy-conditioned adaptation of pretrained affective models.** We formulate adaptation as residual correction of native modality tokens, using encoder hierarchy while reusing the trained encoders, connector, and LLM.
2. **Separate evidence transformations with shared token correction.** Final-relative, depth-specific projections process layer differences; input-dependent aggregation within a shared correction space coordinates their effects on the native tokens.
3. **Cross-benchmark utility and controlled design analysis.** We evaluate cross-benchmark transfer on three backbones and recognition under reduced inputs, then examine the roles of depth conditioning, shared correction, and update direction through controlled comparisons.

2 Related Work

Affective multimodal language models. Affective MLLMs develop task-specialized interfaces through emotion instruction tuning (Cheng et al., 2024), descriptive supervision and pre-fusion (Lian et al., 2025), multiview perception (Peng et al., 2026), and reward-based reasoning (Zhao et al., 2025a). LayerRefit addresses their subsequent adaptation: incorporating encoder-depth evidence while retaining pretrained weights and native token sequences.

Encoder hierarchy and multimodal integration. Multi-layer visual integration spans fusion at the LLM input and within its layers (Lin et al., 2025). At the input, *Depth Tokenization* introduces additional tokens derived from intermediate states (Yao et al., 2024; Chen et al., 2024a), while *Depth Fusion* combines depth features at shared positions

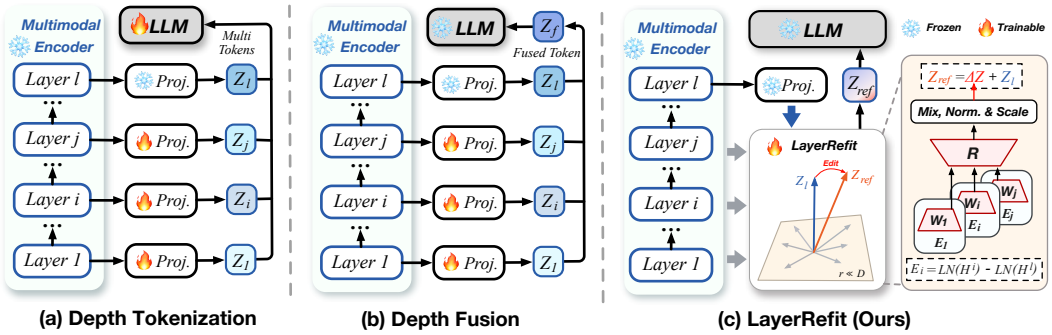


Figure 2: **Encoder-depth integration and evaluated training configurations.** (a) Depth Tokenization adds projected depth tokens. (b) Depth Fusion combines projected depth features at shared token positions. (c) LayerRefit conditions a residual correction of native tokens on final-relative evidence. Training markers show the configurations evaluated here: (a) uses LLM-LoRA, while (b,c) keep the LLM frozen. Evidence projections W_i are depth-specific; correction projection R is shared within each modality and restricts updates to a common low-dimensional subspace.

(Cao et al., 2024; Li et al., 2025; Lin et al., 2026). Internal integration injects encoder features into multiple LLM layers (Shukor et al., 2023; Chen et al., 2024b; Qwen Team, 2025).

Representation-level adaptation. ReFT/LoReFT (Wu et al., 2024) and multimodal methods such as MRT (Liu et al., 2025) and MoReS (Bi et al., 2025) adapt frozen models through representation edits. LayerRefit studies how encoder hierarchy can condition such edits at an established multimodal interface. Intermediate-to-final differences supply corrective evidence; depth-specific transformations and a shared output projection convert this evidence into residual updates of native post-connector tokens.

3 LayerRefit

LayerRefit adds hierarchy-conditioned residual corrections to native modality tokens at the frozen model's post-connector interface. The correction branch combines depth-specific evidence transformations with a shared output projection.

We describe one modality; branch parameters are independent across modalities. Given the frozen connector C and the backbone's original feature selection $\mathcal{H}^{\text{native}}$,

$$Z^{\text{native}} = C(\mathcal{H}^{\text{native}}) \in \mathbb{R}^{T\times D}, \tag{1}$$

$$Z^{\text{refined}} = Z^{\text{native}} + \Delta Z. \tag{2}$$

Figure 2(c) denotes these tokens by Z_l and Z_{ref}, respectively.

3.1 Final-Relative Evidence Projection

After fixed preprocessing aligns token positions, let $H^i \in \mathbb{R}^{T^e\times d}$ denote the state at depth i, with final depth L. For a nonempty set of intermediate depths $\mathcal{S} \subseteq \{1,\dots,L-1\}$, the final state serves as a common reference. Each layer's difference from this reference, computed after normalization, is transformed by a depth-specific evidence projection:

$$E_i = \text{LN}^e(H^i) - \text{LN}^e(H^L), \tag{3}$$

$$Q_i = \text{Pool}_T[\phi(E_i W_i)]. \tag{4}$$

The learned LayerNorm is shared across depths, while each depth has a bias-free evidence projection $W_i \in \mathbb{R}^{d\times r}$. Here ϕ is GELU and Pool_T performs fixed adaptive average pooling, yielding $Q_i \in \mathbb{R}^{T\times r}$.

3.2 Conditional Refinement in a Shared Subspace

A shared, bias-free correction projection $R \in \mathbb{R}^{r \times D}$ maps layer-specific evidence to candidate token updates. An input-conditioned selector weights their contributions:

$$V_i = Q_i R, \qquad w_i = \mathrm{softmax}_{i \in \mathcal{S}}\big(g_i(Z^{\mathrm{native}}, Q_i)\big). \tag{5}$$

The score compares learned summaries of native tokens and projected evidence through cosine similarity, depth biases, and temperature scaling.

We normalize candidates before mixing and scale the aggregate relative to native tokens. Define $s_\epsilon(X) = \sqrt{\mathrm{mean}(X^2) + \epsilon^2}$ and $\mathcal{N}_\epsilon(X) = X / s_\epsilon(X)$, where $\epsilon > 0$ and the mean covers tokens and channels within each sample. Then

$$\begin{aligned} \overline{V} &= \sum_{i \in \mathcal{S}} w_i \mathcal{N}_\epsilon(V_i), \\ \Delta Z &= \rho \tanh(\beta)\, s_\epsilon(Z^{\mathrm{native}}) \mathcal{N}_\epsilon(\overline{V}). \end{aligned} \tag{6}$$

The amplitude β is shared across inputs; depth weights and correction directions are input-dependent. The budget ρ bounds the update by $\mathrm{RMS}(\Delta Z) \le \rho s_\epsilon(Z^{\mathrm{native}})$.

Scalar mixing and RMS normalization preserve the shared correction subspace:

$$\Delta Z[t, :] \in \mathcal{U} = \mathrm{rowspan}(R), \qquad \dim(\mathcal{U}) \le r. \tag{7}$$

Refined tokens lie in $Z^{\mathrm{native}}[t, :] + \mathcal{U}$. The learned projection supplies common output directions, while projected evidence and mixing weights determine each input's update within this space. Additional depth-specific projections thus expand evidence access without increasing the dimension bound r.

3.3 Training from the Native Interface

Randomly initialized evidence and correction projections with $\beta = 0$ yield zero initial correction. Task supervision then trains the refinement branch through the frozen LLM. Restoring the modality index m, we optimize

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathrm{task}} + \lambda_\Delta \mathcal{L}_{\mathrm{reg}}, \\ \mathcal{L}_{\mathrm{reg}} &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{\mathrm{mean}[(\Delta Z_m)^2]}{\max(\mathrm{mean}[(Z_m^{\mathrm{native}})^2], \eta)}. \end{aligned} \tag{8}$$

Here $\mathcal{M}$ contains active modalities, $\lambda_\Delta \ge 0$ controls regularization strength, and $\eta > 0$ stabilizes the denominator. The task loss supervises reference answers through the existing language-model head; Appendix A.4 specifies the training objective.

Training updates only the evidence and correction projections, normalization, selector, and amplitude parameters. Depth sets are selected on source validation before downstream evaluation.

4 Experiments

4.1 Experimental Setup and Protocols

Backbone selection. We evaluate AffectGPT (Lian et al., 2025), Emotion-LLaMAv2 (Peng et al., 2026), and Light-MER (Zheng et al., 2026). The first two support detailed adaptation comparisons; Light-MER tests applicability to a smaller multimodal backbone.

Source adaptation. Each backbone receives its own source-trained branch. Both primary backbones use post-connector refinement, with 0.592M and 0.474M added parameters, respectively. Source supervision comes from a class-balanced MER-Caption+ pool of 864 examples, with training and validation subsets specified per backbone. Hyperparameters are selected on source validation; target labels are excluded from adaptation and selection in the cross-benchmark track. The source cohort has no byte-identical audio/video overlap with the audited MER2023, MER2024, MELD, and IEMOCAP targets, but belongs to AffectGPT's original training distribution.

Comparison routes. **Native** uses the pretrained backbone with the same inference and scoring pipeline. **Final-layer-only refinement ($\mathbf{LR}_{\text{final}}$)** matches LayerRefit's architecture and budget, feeding every evidence projection the same normalized final encoder state. **DF** directly fuses multi-depth features before the LLM. $\mathbf{DT}_{\text{LoRA-64}}$ appends projected tokens from multiple encoder layers as additional LLM prompts and trains LoRA to process the expanded sequence. For each adaptation route, backbone, and training seed, we train once with A+V+T and evaluate six input configurations without retraining (Table 2). Counts include all added parameters; only the supplied modalities change at inference. Fixed-checkpoint on/off interventions are separate from retrained controls.

Evaluation tracks. Frozen cross-benchmark evaluation follows MER-UniBench (Lian et al., 2025), using WAF for sentiment, HIT for basic emotion, and emotion-wheel F-score (F_s) for fine-grained understanding. Avg-9 is the descriptive arithmetic mean of the nine displayed scores; Table 2 reports native-generation HIT (%) on MER2024. Controlled recognition uses a fixed six-class candidate-label readout with Macro-F1 and NLL; the latter measures true-label probability assignment. Protocols and backbone details appear in Appendix A; refinement hyperparameters and training settings are listed in Appendix A.4.

4.2 Main Results and Generalization

Table 1 evaluates a separately source-trained and then frozen branch for each of three backbones on nine benchmarks. Comparisons with the corresponding Native rows measure refinement gains; other published systems provide context under their original setups.

Table 1: **Cross-benchmark results on MER-UniBench (%).** Metrics are WAF for sentiment, HIT for basic emotion, and emotion-wheel F-score (F_s) for fine-grained understanding. Avg-9 averages the nine scores; parentheses give its change from the corresponding backbone. Indented rows add LayerRefit; other MLLMs provide published reference scores. Bold and underlining mark the best and second-best scores per column.

Model	Sentiment (WAF)				Basic emotion (HIT)				Fine-grained (F_s)	Mean
	MOSI	MOSEI	SIMS	SIMSv2	MER23	MER24	MELD	IEMOCAP	OV-MERD+	Avg-9
R1-Omni-0.5B (2025a)	58.02	56.48	71.82	68.58	64.17	67.43	43.20	51.58	55.24	59.61
Emotion-LLaMA (2024)	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
OPPO (2026)	86.50	84.63	86.22	**88.26**	**87.73**	**90.34**	**64.06**	73.60	**67.16**	**80.94**
AffectAgent (AGPT) (2026)	82.73	82.67	89.43	87.33	80.69	80.66	59.07	63.39	65.13	76.79
HumanOmni-7B (2025b)	7.37	7.33	43.87	35.35	63.36	76.64	37.05	52.52	59.03	42.50
HumanOmniV2-9B (2025)	76.56	75.44	83.58	81.65	68.59	57.41	42.74	60.95	61.62	67.62
Qwen2.5-Omni-7B (2025)	78.86	61.84	75.44	76.67	72.84	80.44	43.72	61.53	53.13	67.16
AffectGPT (2025)	81.30	80.90	88.49	86.18	78.54	78.80	55.65	60.54	62.52	74.77
↪ + **LayerRefit (ours)**	83.86	82.41	**90.92**	87.68	81.81	81.97	56.74	64.19	62.40	76.89 (+2.12)
Emotion-LLaMAv2 (2026)	86.28	**87.69**	87.72	86.02	77.28	86.90	51.32	84.05	62.90	78.91
↪ + **LayerRefit (ours)**	**87.84**	87.53	89.15	86.97	79.46	88.17	52.09	**86.21**	63.72	80.13 (+1.22)
Light-MER (2026)	80.04	77.23	89.05	87.74	75.66	80.04	57.02	61.46	63.28	74.61
↪ + **LayerRefit (ours)**	80.04	77.10	90.21	87.52	78.38	83.84	59.45	65.27	63.87	76.19 (+1.57)

Colors in LayerRefit rows indicate gains (green), losses (red), or no change (black) relative to the corresponding backbone. Averages and changes are computed before rounding. R1-Omni scores are taken from Han et al. (2026), Table 2; HumanOmni, HumanOmniV2, and Qwen2.5-Omni scores are taken from Peng et al. (2026), Table IV. AGPT denotes AffectGPT.

Consistent gains across backbones. LayerRefit raises Avg-9 by 2.12, 1.22, and 1.57 points on AffectGPT, Emotion-LLaMAv2, and Light-MER, respectively (Table 1). Each backbone uses its own source-trained branch, extending the gains across three distinct pretrained multimodal architectures.

Generalization across downstream tasks. LayerRefit improves 22 of 27 evaluations, including all 12 basic-emotion settings across four benchmarks and three backbones. Sentiment improves in 8 of 12 settings, and fine-grained understanding improves on two of the three backbones. Basic-emotion recognition shows the most consistent benefit from source-trained refinement, with gains extending to broader affective tasks.

4.3 Inference across Input Configurations

To assess whether gains persist across inference-time input compositions, Table 2 evaluates six configurations on AffectGPT and Emotion-LLaMAv2. Each adaptation route is trained with A+V+T under shared source supervision and evaluated with unchanged weights as the supplied modalities vary. Native provides the reference, and performance is measured by MER2024 native-generation HIT.

Table 2: **Inference across input configurations on MER2024 (HIT, %).** $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$ provides a larger-budget LoRA comparison (Hu et al., 2022). $\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$ and H-LoReFT (our hierarchy-aware adaptation of LoReFT; Wu et al., 2024) have added parameter counts comparable to LayerRefit. Each A+V+T-trained model is evaluated without retraining. Δ_{avg} averages changes from Native over six configurations; parameter percentages are relative to $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$. Colors indicate changes from Native; bold and underlining mark the best and second-best performance within each backbone.

Method	Params (M)	A	V	A+V	A+T	V+T	A+V+T	Δ_{avg}
AffectGPT (2025)		59.48	56.10	72.42	73.41	75.29	78.80	0.00
↪ $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$	+183.800 (100.00%)	62.19	58.67	**77.92**	**78.93**	**77.21**	**84.40**	**+3.97**
↪ DF	+22.100 (12.02%)	58.45	57.27	72.35	70.64	69.73	76.18	-1.81
↪ $\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$	+0.588 (0.32%)	59.26	56.54	73.72	74.98	75.29	79.37	+0.61
↪ H-LoReFT	+0.592 (0.32%)	58.39	55.70	71.49	70.82	70.26	75.93	-2.15
↪ $\mathrm{LR}_{\mathrm{final}}$	+0.592 (0.32%)	59.60	54.65	71.06	73.23	73.20	77.18	-1.10
↪ **LayerRefit**	+0.592 (0.32%)	**62.92**	**59.45**	76.81	77.48	75.25	81.97	+3.06
Emotion-LLaMAv2 (2026)		60.78	55.65	78.92	72.67	67.46	86.90	0.00
↪ $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$	+176.730 (100.00%)	**63.87**	**58.12**	**82.97**	72.82	66.24	85.75	+1.23
↪ DF	+16.822 (9.52%)	58.69	56.74	77.67	70.69	68.48	84.40	-0.95
↪ $\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$	+0.473 (0.27%)	61.01	55.92	79.69	72.03	66.79	85.37	-0.26
↪ H-LoReFT	+0.474 (0.27%)	58.43	54.76	77.08	70.92	66.47	86.41	-1.39
↪ $\mathrm{LR}_{\mathrm{final}}$	+0.474 (0.27%)	57.93	53.45	76.82	73.08	66.47	83.65	-1.83
↪ **LayerRefit**	+0.474 (0.27%)	61.92	57.10	82.10	**74.48**	**71.05**	**88.17**	**+2.07**

A, V, and T denote audio, vision, and text. DF denotes direct depth fusion; DT denotes depth tokenization with LLM-LoRA. Final-layer-only refinement ($\mathrm{LR}_{\mathrm{final}}$) retains all evidence projections, the selector, and the shared correction projection at the same parameter budget, using final-layer features throughout. Pretrained weights remain fixed; DT additionally trains LoRA updates to the LLM.

Gains across inputs without retraining. LayerRefit improves 11 of 12 configurations, with mean HIT gains of 3.06 and 2.07 points on AffectGPT and Emotion-LLaMAv2. These gains persist when inputs are restricted to individual modality streams or subsets of the training modalities. $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$ achieves the larger mean gain on AffectGPT, whereas LayerRefit leads on Emotion-LLaMAv2. LayerRefit's added parameter counts are 0.32% and 0.27% of those of $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$ on the respective backbones. With comparable parameter budgets, LayerRefit outperforms $\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$ in 11 of 12 configurations, with average margins of 2.45 and 2.34 HIT points, respectively.

Intermediate evidence and native-token correction. At matched budgets, LayerRefit exceeds $\mathrm{LR}_{\mathrm{final}}$ in all 12 input configurations, by 4.16 and 3.90 average HIT points on the two backbones. $\mathrm{LR}_{\mathrm{final}}$ itself improves only two configurations over Native. With structure and parameter budget matched, these results support intermediate-state conditioning. LayerRefit also exceeds DF and H-LoReFT in all 12 configurations, with H-LoReFT having comparable added parameter counts. Together with the final-layer-only control, these results support both access to intermediate evidence and its formulation as a native-token correction. The following ablations examine how evidence conditioning and the correction structure contribute to these gains.

4.4 Design and Correction Analysis

To identify the design choices that make hierarchical native-token correction effective, we compare evidence conditioning, evidence projection and token correction, depth selection and aggregation,

and edit location (Table 3). Figure 3 complements these ablations with oracle correction geometry (a), condition-dependent depth utility (b), and norm-matched directional interventions at a fixed checkpoint (c). We evaluate the variants on MER2024 using AffectGPT with A+V+T inputs and a controlled six-class readout (angry, happy, neutral, sad, surprise, and worried). Each variant alters one design axis under a common data split, initialization protocol, and optimization procedure. Matched-budget comparisons control added parameter counts; the remaining variants compare architectures at their reported budgets.

Table 3: **Core design ablation results on MER2024.** AffectGPT is evaluated with A+V+T inputs using a six-class candidate-label readout. Macro-F1 (↑) measures class-balanced recognition; NLL (↓) measures true-label negative log-likelihood. Parentheses give the Macro-F1 change from full LayerRefit. Params reports total added parameters in millions.

Variant	Params (M)	Macro-F1 ↑ (Δ)	NLL ↓
AffectGPT (2025)	0	56.94 (−4.93)	1.8652
↪ **LayerRefit (ours)**	0.592	**61.87** (0.00)	1.608
A. Evidence conditioning			
Absolute intermediate	0.592	55.62 (−6.25)	1.713
Final-layer only ($\mathrm{LR}_{\mathrm{final}}$)	0.592	56.67 (−5.20)	1.689
B. Evidence projection and token correction			
Shared evidence projection	0.232	57.49 (−4.38)	1.654
Independent correction projections	1.223	59.46 (−2.41)	1.582
C. Depth selection and aggregation			
Uniform aggregation	0.403	56.73 (−5.14)	1.646
Global Top-K depths	0.592	60.08 (−1.79)	1.639
D. Edit location			
Multi-layer LLM editing	0.561	59.15 (−2.72)	**1.551**
Pre-connector editing	0.161	59.68 (−2.19)	1.631

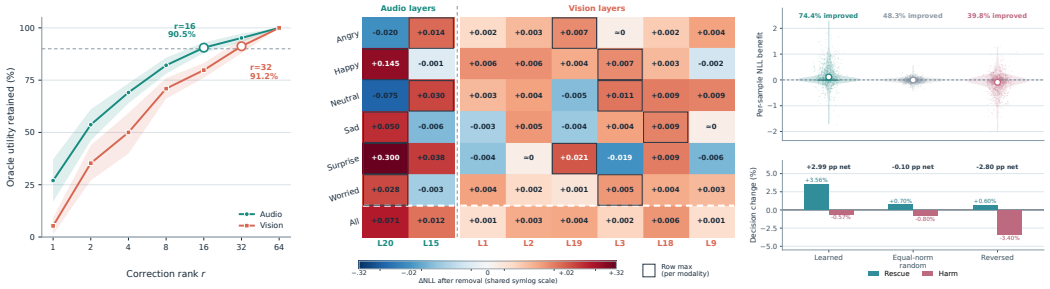
Final-relative evidence. LayerRefit improves Macro-F1 over Native by 4.93 points and lowers NLL from 1.8652 to 1.608. At the same 0.592M budget, absolute-intermediate and final-layer-only refinement trail LayerRefit by 6.25 and 5.20 points, respectively. The absolute control retains the normalization, evidence projections, and selector. Both controls lower NLL but trail Native in Macro-F1. The matched-budget comparisons favor both intermediate-depth conditioning and its expression relative to the final encoder state.

Depth-specific evidence projection and shared correction. Sharing the evidence projection lowers Macro-F1 by 4.38 points (0.232M). Retaining depth-specific transformations helps accommodate differences among encoder representations. Independent correction projections use 1.223M parameters yet trail LayerRefit by 2.41 points, despite lower NLL (1.582 versus 1.608). For class-balanced recognition, this contrast favors coordinating layer-specific evidence through a shared output mapping: separate transformations accommodate depth differences, while their input-dependent combination determines the update within the shared correction space. The separate oracle diagnostic (Figure 3(a)) retains 90.5% and 91.2% of rank-64 NLL improvement at audio rank 16 and vision rank 32, respectively, providing geometric motivation for compact correction spaces.

Depth selection and conditional aggregation. Joint source-validation selection exceeds Global Top-K by 1.79 Macro-F1 points at matched depth count and budget; the smaller uniform-aggregation variant trails LayerRefit by 5.14 points. Removing audio layer 20 raises NLL by 0.300 for Surprise but lowers it by 0.075 for Neutral (Figure 3(b)): the same contribution helps one emotion and interferes with another. This conditional utility extends the probing observations in Figure 1 to learned corrections, motivating joint depth selection followed by input-dependent weighting of the selected evidence.

Edit location. Post-connector refinement reaches 61.87 Macro-F1, versus 59.68 for pre-connector refinement and 59.15 for multi-layer LLM editing. The latter has the lowest NLL (1.551), yet trails LayerRefit in Macro-F1. Across Table 3, all adaptation variants lower NLL relative to Native, while

three also lower Macro-F1. The combined evidence favors a specific allocation of capacity for class-balanced recognition: separate transformations of final-relative features, followed by token correction through a common output space.



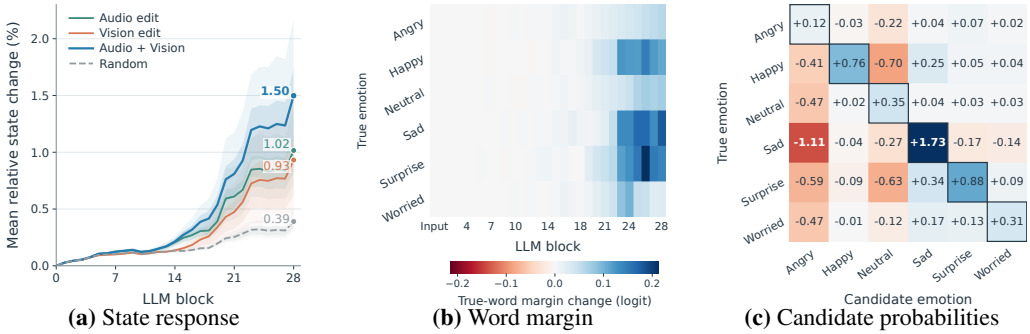
(a) Compact geometry **(b)** Conditional depth utility **(c)** Directional validity

Figure 3: **Geometry, conditional depth utility, and directional validity of LayerRefit.** (a) Separate held-out oracle diagnostic: NLL improvement retained relative to rank 64. (b) Class-conditioned change in true-label NLL after removing one depth contribution and renormalizing the remainder, $\Delta\mathrm{NLL} = \mathrm{NLL}_{-i} - \mathrm{NLL}_{\mathrm{full}}$. Positive values indicate harmful removal; outlines mark row maxima per modality, and the final row aggregates classes. (c) Learned, equal-norm random, and sign-reversed joint corrections. Upper: sample-level $\mathrm{NLL}_{\mathrm{off}} - \mathrm{NLL}_{\mathrm{edit}}$; percentages report the fraction with positive benefit. Lower: rescue (wrong to correct) and harm (correct to wrong) rates; net change is rescue minus harm in percentage points.

Task utility depends on correction direction. At matched norms, learned corrections improve true-label likelihood for 74.4% of examples, versus 48.3% for random directions and 39.8% for reversal (Figure 3(c)). Net decision changes are $+2.99$, -0.10, and -2.80 percentage points, respectively. With magnitude controlled, the change from improvement to harm follows the correction direction. These interventions complement Table 3 by testing the task value of the resulting update.

4.5 Propagation and Conditional Benefits

To characterize how corrections affect predictions, we analyze the pretrained, frozen AffectGPT backbone with trained LayerRefit branches. Figure 4 follows input-token edits to the final context position before answer generation; Figure 5 examines their benefits across emotions and Native prediction difficulty.

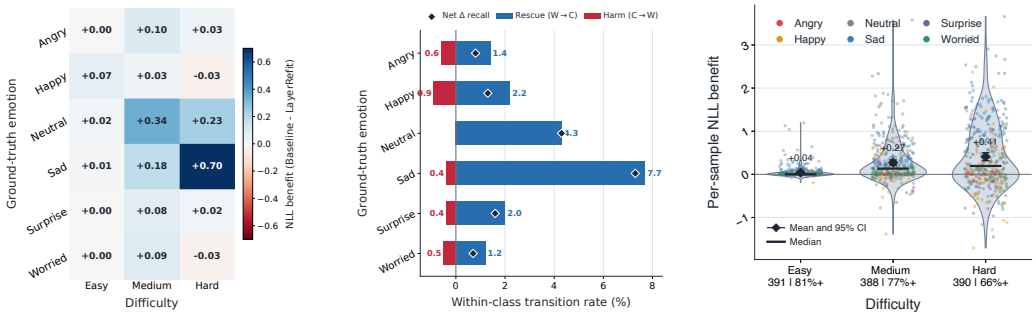


(a) State response **(b)** Word margin **(c)** Candidate probabilities

Figure 4: **Propagation to the answer readout in frozen AffectGPT.** Panels compare refinement on versus off at the answer context position. (a) Relative hidden-state change, $100\|h_{\mathrm{edit}} - h_{\mathrm{off}}\|/\|h_{\mathrm{off}}\|$ (%), across blocks 0 to 28 for audio, vision, joint, and equal-norm random edits. Lines show sample means; shading shows sample variation. (b) Joint-edit change in the true-word logit margin over the strongest competitor, read through the fixed output head at each block and averaged by true class. (c) Final-readout probability changes after six-candidate softmax, in percentage points. Rows are true classes, columns candidates; outlines mark correct candidates.

Propagation through the frozen backbone. Input corrections propagate to the answer context position, with the largest state responses in later blocks (Figure 4(a)). Joint edits yield larger late-layer responses than either modality alone or equal-norm random edits. At the same position, late-layer word-margin changes differ across emotions (b), connecting the state response to changes in the emotion-word readout of the frozen backbone.

Changes in emotion discrimination. At the final readout, mean true-class probability increases for all six classes (Figure 4(c)). Sad shows the largest increase (+1.73 percentage points), accompanied by a 1.11-point decrease in Angry probability on Sad examples. This redistribution favors the correct emotion over a competing class. Sad also has the largest net recall gain, with 7.7% rescues and 0.4% harms (Figure 5(b)), linking stronger class preference to improved recognition.



(a) Emotion and difficulty benefit **(b)** Rescue and harm **(c)** Per-sample benefit

Figure 5: **Refinement benefits across emotions and prediction difficulty.** Benefit is $\mathrm{NLL}_{\mathrm{Native}} - \mathrm{NLL}_{\mathrm{LayerRefit}}$; difficulty is defined from Native true-label NLL. (a) Mean benefit by true emotion and difficulty. (b) Within-class rescue (wrong to correct) and harm (correct to wrong) rates; diamonds show their difference, the recall change in percentage points. (c) Sample-level benefits with group means and their 95% confidence intervals, and medians. Labels give sample counts and the fraction of examples with positive NLL benefit.

Refinement addresses condition-dependent recognition deficits. Figure 5 extends the conditional analysis from individual depth contributions (Figure 3(b)) to complete corrections, grouping their benefits by emotion and Native prediction difficulty. Difficulty measures Native true-label NLL, so hard examples have weaker true-class support in the original readout. Mean NLL benefit rises from 0.04 for easy inputs to 0.41 for hard inputs, while improvement coverage falls from 81% to 66% (Figure 5(c)). Larger absolute reductions thus coexist with less widespread benefit. The pattern also depends on emotion: hard-group benefit is 0.70 for Sad and -0.03 for Happy (a). Happy's negative hard-group benefit occurs alongside an overall increase in mean true-class probability (+0.76 points in Figure 4(c)), showing how class-level readout gains can coexist with uneven benefits across difficulty groups. Refinement therefore partly compensates for weak native predictions in selected groups defined by emotion and difficulty, with benefit magnitude and coverage capturing distinct aspects of this improvement.

5 Conclusion

LayerRefit adapts pretrained affective MLLMs by converting final-relative encoder evidence into residual corrections of native modality tokens. Cross-benchmark and reduced-input evaluations demonstrate the utility of these corrections while backbone weights and token counts remain fixed. Ablations favor depth-specific evidence transformations with a shared correction mapping, and norm-matched interventions connect recognition gains to the learned update direction. Together, these results support a design principle for interface adaptation: process heterogeneous depth evidence through distinct transformations, then coordinate its input-dependent contributions within a common correction space.

AI USE STATEMENT

We used OpenAI Codex to assist with manuscript wording, translation, and organization, literature identification, feedback on experimental framing, interpretation of existing results, and implementation and testing of the released code. The authors reviewed the AI-assisted material and take responsibility for the final text, analyses, and claims.

ETHICS STATEMENT

This work evaluates existing audiovisual affective benchmarks and pretrained models. Speech and facial recordings can contain sensitive personal information, while emotion labels depend on context and annotation conventions. Our results characterize benchmark performance and do not establish suitability for individual-level or high-stakes emotion assessment.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the refinement method and evaluation design. The appendix provides backbone configurations, training settings, encoder-probing protocols, correction-capacity studies, and additional answer-context response analyses.

REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15230–15250, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.739. URL https://aclanthology.org/2025.acl-long.739/.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu. MMFuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. *arXiv preprint arXiv:2410.11829*, 2024.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26540–26550, 2024a.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6677–6695. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.339.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, Changyi Liu, Dewen Fan, Huihui Xiao, Jiahong Wu, Fan Yang, Size Li, and Di Zhang. EVLM: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*, 2024b.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Zhiyuan Han, Beier Zhu, Wenwen Tong, Pengyang Shao, Peipei Song, Xinyi Wang, Jiangnan Chen, Lewei Lu, and Xun Yang. Omni-perception policy optimization for multimodal emotion reasoning. *arXiv preprint arXiv:2606.25325*, 2026.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Patrik N. Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814, 2003. doi: 10.1037/0033-2909.129.5.770.

Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, Bin Li, and Xiangyang Xue. Instruction-guided fusion of multi-layer visual features in large vision-language models. *arXiv preprint arXiv:2501.08443*, 2025.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. *arXiv preprint arXiv:2304.08981*, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2404.17113*, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Chenchen Lin, Sanbao Su, Rachel Luo, Yuxiao Chen, Yan Wang, Marco Pavone, and Fei Miao. TGIF: Text-guided layer fusion mitigates hallucination in multimodal LLMs. In *Conference on Language Modeling*, 2026.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal LLMs: Methods, analysis, and best practices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, and Cheng Han. Re-imagining multimodal instruction tuning: A representation view. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dcf88cbc8d01ce7309b83d0ebaeb9d29-Abstract-Conference.html.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with VoxParadox. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. Emotion-LLaMAv2 and MMEVerse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alexandra Saliba, Yuanchao Li, Ramon Sanabria, and Catherine Lai. Layer-wise analysis of self-supervised acoustic word embeddings: A study on speech emotion recognition. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 590–594, 2024. doi: 10.1109/ICASSPW62465.2024.10627629.

Klaus R. Scherer and Heiner Ellgring. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1):158–171, 2007. doi: 10.1037/1528-3542.7.1.158.

Mustafa Shukor, Corentin Dancette, and Matthieu Cord. eP-ALM: Efficient perceptual augmentation of language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22069, 2023. doi: 10.1109/ICCV51070.2023.02016. URL `https://arxiv.org/abs/2303.11403`.

Zeheng Wang, Zitong Yu, Yijie Zhu, Bo Zhao, Haochen Liang, Taorui Wang, Wei Xia, Jiayu Zhang, Zhishu Liu, Hui Ma, Fei Ma, and Qi Tian. AffectAgent: Collaborative multi-agent reasoning for retrieval-augmented multimodal emotion recognition. *arXiv preprint arXiv:2604.12735*, 2026.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, Yuxin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for MLLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.343.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025a.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan Chen, Xihan Wei, and Liefeng Bo. HumanOmni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025b.

Kaiwen Zheng, Junchen Fu, Wenhao Deng, Hu Han, Joemon M. Jose, and Xuri Ge. Do we really need multimodal emotion language models larger than 1b parameters? *arXiv preprint arXiv:2607.12787*, 2026.

A Experimental Protocols and Implementation

The supplementary analyses follow the same progression as the main paper: from the distribution of affective evidence across encoder depth to the capacity of token corrections and their propagation through a frozen LLM. This section specifies the evaluation tracks and backbone configurations. Appendices B and C examine encoder readouts and conditional feature structure; Appendices D and E study correction capacity and answer-context responses.

A.1 Training and Evaluation Tracks

Each backbone receives its own source-trained LayerRefit branch. Model selection uses source validation, after which the selected checkpoint is held fixed for cross-benchmark evaluation. The class-balanced MER-Caption+ source pool contains 864 records, split into 696 training, 84 validation, and 84 source-test records. All three backbones use this split; Table 8 gives their optimization settings.

For the input-configuration comparison in Table 2, each adaptation route uses one A+V+T-trained checkpoint per backbone and training seed. The six configurations change only the modalities supplied at inference. Cross-benchmark evaluation retains the native WAF, HIT, and F_s definitions. For evaluation, the structural ablations in Table 3 use a six-label candidate readout on MER2024 and report Macro-F1 and candidate-normalized true-label NLL. This readout evaluates emotion discrimination in a fixed candidate set; branch training uses the autoregressive objective defined in Appendix A.4.

Table 4: Evaluation tracks and their selection protocols. Encoder diagnostics, source-data oracle interventions, and downstream recognition use distinct readouts and sample sets.

Study	Samples and readout	Selection and comparison
Cross-benchmark transfer	Nine tasks; WAF, HIT, or F_s	Source-validation selection; fixed checkpoints on target benchmarks.
Input configurations	MER2024 native-generation HIT	One A+V+T checkpoint per backbone, route, and seed; six inference-time input combinations.
Structural ablations	MER2024 A+V+T; six-label Macro-F1 and NLL	One design axis varies at a time; counts cover all added parameters.
Encoder diagnostics	CREMA-D emotion and intensity conditions	Frozen encoder features and probe readouts.
Cross-dataset probes	Six datasets; internal validation accuracy	Cached features; approximately 9:1 internal split; peak validation accuracy per layer.
Local oracle geometry	Source records; task-loss interventions	Disjoint basis-fitting and evaluation sets; true-label gradients specify diagnostic directions.

A.2 Backbone Interfaces and Refinement Configuration

LayerRefit operates once on native modality tokens after the connector and before the first LLM block. The encoders, connector, and LLM remain frozen. Table 5 lists the encoders used by each backbone. The Light-MER encoder diagnostics use full video frames, whereas its native visual pathway uses face crops.

Table 5: Frozen encoders and visual inputs. The full-frame Light-MER setting applies to the encoder diagnostics in this appendix.

Backbone	Audio encoder	Visual encoder	Visual input
AffectGPT	Chinese HuBERT-large	CLIP ViT-L/14	Native preprocessing.
Emotion-LLaMAv2	Whisper-large-v3	EVA-ViT-G	Native preprocessing.
Light-MER	Chinese HuBERT-base	CLIP ViT-B/16	Full frames for encoder diagnostics; face crops in the native pathway.

The supplementary figures use the main-text notation: E_i is final-relative evidence, W_i is a depth-specific evidence projection, and R is the shared correction projection within a modality. Thus $Q_i = \mathrm{Pool}_T[\phi(E_i W_i)]$ and $V_i = Q_i R$. Input-conditioned weights combine the candidates, and the normalized aggregate yields the residual update ΔZ of Z^{native}. Table 6 gives the selected layer sets and correction ranks.

For each backbone, we compare a small set of predefined depth configurations covering shallow, middle, deep, and mixed-depth selections, excluding the final reference layers. With depth counts, correction ranks, and training budgets held fixed, we select the configuration with the lowest source-validation NLL and retain it across all target benchmarks and inference-time input configurations.

Table 6: Selected encoder depths and correction ranks. K_a and K_v denote the numbers of selected audio and visual depths; r_a and r_v denote the corresponding correction ranks.

Backbone	Audio depths	Visual depths	K_a/K_v	r_a/r_v
AffectGPT	20, 15	1, 2, 19, 3, 18, 9	2/6	16/32
Emotion-LLaMAv2	5, 29, 3	36, 8, 32, 24, 16, 39	3/6	16/8
Light-MER	4, 9	3, 5, 7, 8, 10, 11	2/6	32/32

A.3 Baselines with Comparable Parameter Budgets

To compare ways of using multi-layer information at similar added parameter budgets, we construct $\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$ and H-LoReFT for Table 2. Their low-rank dimensions are chosen for an added parameter budget close to LayerRefit's on each backbone. Both use the same selected encoder layers as LayerRefit (Table 6). Counts include all added projection and editing modules, excluding frozen pretrained parameters.

$\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$. Within each modality, a low-rank projection shared across depths maps intermediate features to additional token blocks. Query and value projections in the first seven LLM blocks receive LoRA adapters (Hu et al., 2022). For AffectGPT and Emotion-LLaMAv2, respectively, the token-projection ranks are 12 and 23, and the Q/V LoRA ranks are 6 and 3.

H-LoReFT. We align and uniformly average final-relative evidence from the selected layers, then use a shared projection to condition the target of the LoReFT subspace edit (Wu et al., 2024). An independent editor applies one correction per modality, at AffectGPT's post-projection token interface and Emotion-LLaMAv2's native pre-projection feature interface. The audio/visual editor ranks are 19/53 and 28/86, respectively. This adaptation retains the learnable transformation of the current representation while adding conditioning on multi-layer encoder evidence.

Table 7 gives the exact added parameter counts. Each baseline differs from the corresponding LayerRefit budget by less than 1%, enabling comparison of the information-integration structures at similar parameter counts.

Table 7: Exact added parameter counts in millions for the budget-comparable baselines and LayerRefit. Table 2 reports these values rounded to three decimal places.

Backbone	$\mathrm{DT}_{\mathrm{LoRA\text{-}small}}$	H-LoReFT	LayerRefit
AffectGPT	0.587776	0.591944	0.592036
Emotion-LLaMAv2	0.473088	0.473586	0.473765

A.4 Optimization Settings

LayerRefit is trained with autoregressive language-model supervision on reference answers. Let x denote the supplied multimodal input, q the instruction, and $y = (y_1, \ldots, y_N)$ the reference answer tokens. With frozen backbone parameters θ and refinement parameters ψ, the task loss in Eq. 8 is

$$\mathcal{L}_{\mathrm{task}} = -\frac{1}{N} \sum_{t=1}^{N} \log p_{\theta,\psi}(y_t \mid y_{<t}, x, q). \tag{9}$$

Teacher forcing supplies the reference answer prefix at each step, and supervision is applied to answer tokens through the existing language-model head over the full vocabulary. Instruction and padding positions are excluded from the loss. Gradients propagate through the frozen LLM to the refinement branch. The six-candidate NLL reported for controlled recognition is an evaluation measure computed after normalization over the candidate labels, distinct from the autoregressive training loss.

Refinement hyperparameters. The audio and visual RMS budgets are $\rho_a = \rho_v = 0.5$, and the relative-update regularization coefficient is $\lambda_\Delta = 10^{-3}$. The selector temperature is fixed at $\tau = 0.5$. Numerical stability constants are $\epsilon_{\mathrm{RMS}} = \epsilon_{\cos} = 10^{-6}$ for RMS normalization and cosine similarity, $\epsilon_{\mathrm{LN}} = 10^{-5}$ for LayerNorm, and $\eta = 10^{-8}$ for the regularization denominator. Here ϵ_{RMS} corresponds to ϵ in Eq. 6. Amplitude parameters and depth biases are initialized to zero and learned during training. The refinement branches use no dropout.

Scores from our training runs in Tables 1–3 are averaged over three runs with seeds 42, 43, and 44. Reported gains and improvement counts are computed from these mean scores. Figure 5(a–c) reports results from a single run with seed 42. Table 8 reports the backbone-specific learning rates and training budgets.

Table 8: Source-training recipes. All three backbones use AdamW and seeds 42, 43, and 44.

Backbone	Learning rate and training budget
AffectGPT	3×10^{-4}; 174 steps per epoch, at most 6 epochs.
Emotion-LLaMAv2	10^{-3}; amplitude parameter learning rate 10^{-2}; 693 steps.
Light-MER	10^{-3}; at most 6 epochs.

For AffectGPT, the depth-specific W_i and selector projections are independently initialized from $\mathcal{N}(0, 0.02^2)$, and the shared correction projection R from $\mathcal{N}(0, 0.01^2)$. Training uses weight decay 10^{-4}, batch size 4, gradient-norm clipping at 1.0, and a constant learning rate without a scheduler. Source-validation NLL selects the checkpoint.

A.5 Inference Efficiency

To complement the parameter comparison in Table 2, Table 9 reports inference latency for Native, LayerRefit, and three comparison routes with AffectGPT. Measurements use an RTX PRO 6000 GPU, FP16 precision, and batch size 1 with A+V+T inputs. All methods process the same 16 media samples after two warm-up runs per method and generate exactly 24 tokens. Timing includes preprocessing and end-to-end inference, excluding model loading. The p50 and p95 values summarize the latency distribution across the measured samples.

Table 9: Inference latency of the compared adaptation methods. Added parameters are reported in millions. Latencies are in seconds; p50 and p95 denote the 50th and 95th percentiles. Overhead is the relative increase in mean latency over Native.

Method	Params (M)	Mean ↓	p50 ↓	p95 ↓	Overhead
Native	0.000	2.797	2.789	3.235	0.00%
LayerRefit (ours)	**0.592**	**2.866**	**2.810**	**3.357**	**+2.47%**
$\mathrm{LR}_{\mathrm{final}}$	0.592	2.865	2.782	3.491	+2.45%
DF	22.100	2.828	2.774	3.124	+1.14%
$\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$	183.800	3.741	3.715	3.961	+33.77%

Under this fixed-length generation protocol, LayerRefit and final-layer-only refinement have similar mean latencies (2.866 and 2.865 s). LayerRefit adds 0.069 s over Native, a relative overhead of 2.47%. The p50 and p95 latencies increase by 0.021 s and 0.122 s over Native, respectively. DF incurs a 1.14% mean overhead, while $\mathrm{DT}_{\mathrm{LoRA\text{-}64}}$ incurs 33.77%. Together with the recognition results in Table 2, these measurements show that LayerRefit's use of hierarchical evidence is compatible with modest inference overhead and a small added parameter budget.

B AFFECTIVE EVIDENCE ACROSS ENCODER DEPTH

Figure 1 motivates refinement from two complementary views: depth preferences across benchmarks and conditional differences within CREMA-D. We extend both views to the three backbone families, first examining emotion and intensity groups and then cross-dataset readout profiles.

B.1 EMOTION- AND INTENSITY-CONDITIONED PROBES

The three backbones share 1,092 CREMA-D clips from 91 actors. Linear probes use five actor-disjoint folds with seed 42 and report joint recall for four emotions at three expression intensities. Each heatmap gives percentage-point changes relative to the final encoder layer of the same backbone and modality. Outlines mark a layer attaining the highest recall within a row. The shared color range is $[-30, 30]$ percentage points; L/M/H, also written LO/MD/HI, denote low, medium, and high expression intensity.

AffectGPT. Figures 6 and 7 show all 24 HuBERT and ViT layers. They expand the representative depths in Figure 1(c) into complete profiles. Peak depths vary across emotion and intensity groups. In audio, for example, the largest gain for Joy occurs at different depths for low, medium, and high intensity. This conditional structure motivates access to several depths whose contributions can vary with the input.

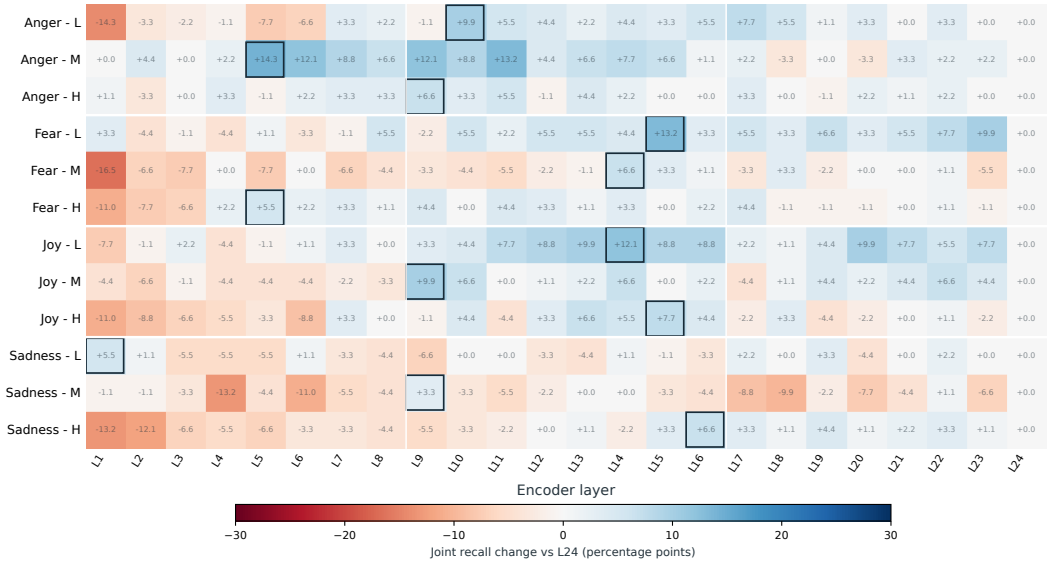


Figure 6: AffectGPT HuBERT audio probes across L1–L24. Cells report joint recall changes relative to L24 in percentage points, using the shared actor-disjoint protocol. Outlines mark row maxima.

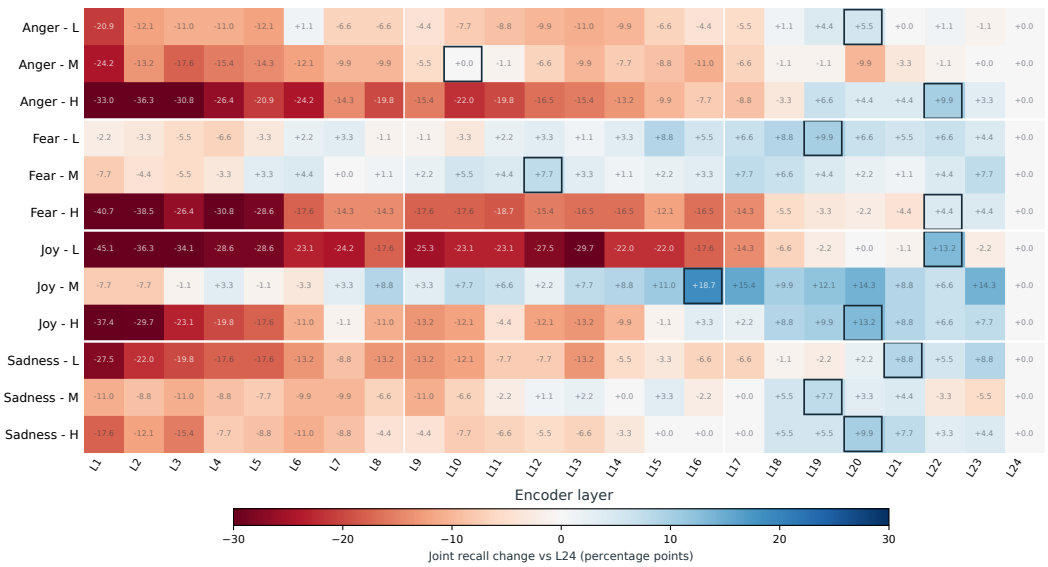


Figure 7: AffectGPT ViT visual probes across L1–L24. Samples, actor folds, seed, and color scale match Figure 6; the reference is the visual encoder's L24.

Emotion-LLaMAv2. Figures 8 and 9 extend the same samples and probe protocol to Whisper-large-v3 and EVA-ViT-G. Audio and visual changes are measured against their respective final layers, L32 and L40.

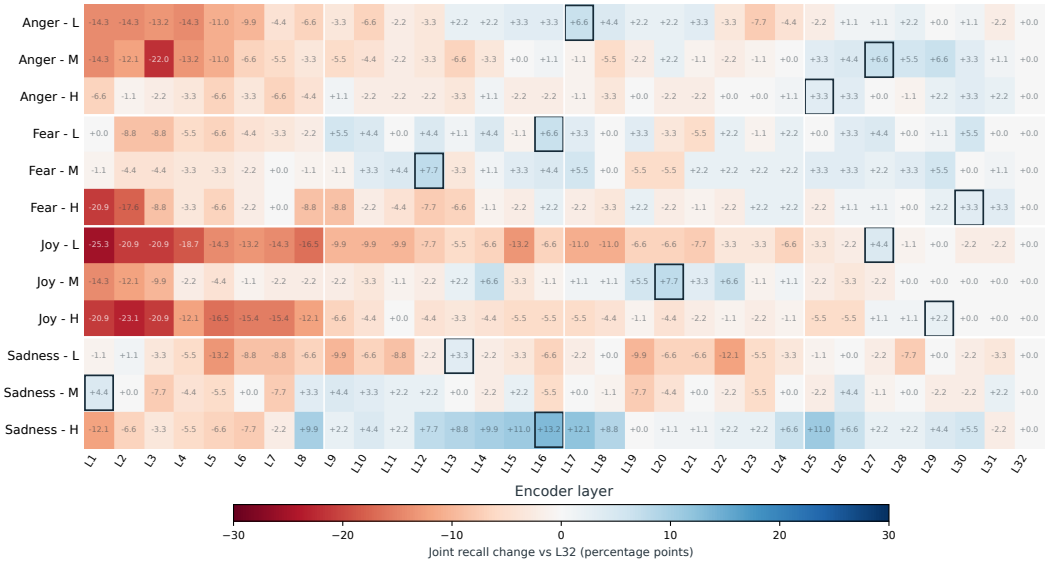


Figure 8: Emotion-LLaMAv2 Whisper-large-v3 audio probes across L1–L32. Each cell is relative to audio L32, with the same actor folds, probe protocol, and color scale as the AffectGPT analysis.

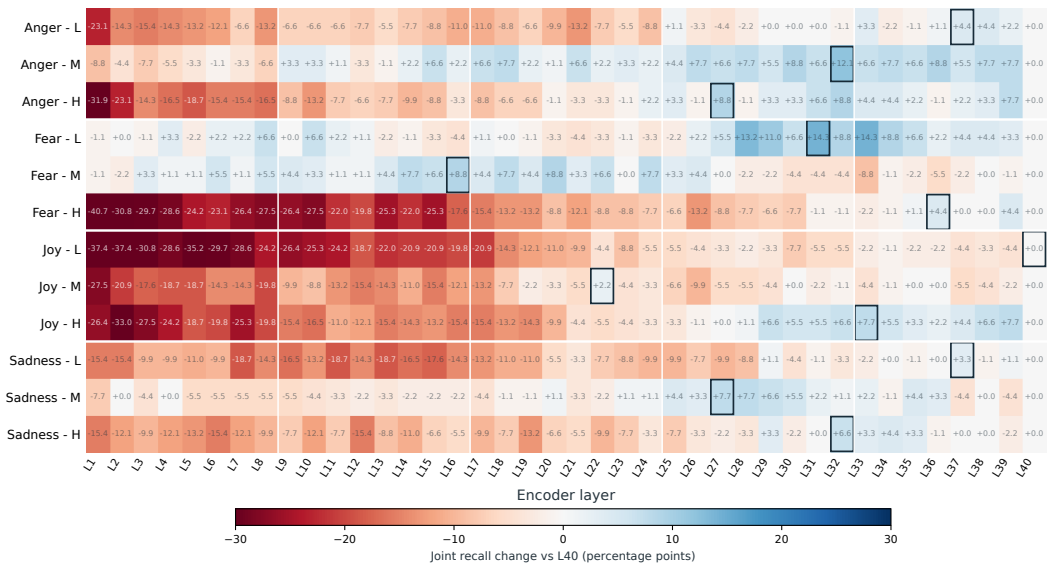


Figure 9: Emotion-LLaMAv2 EVA-ViT-G visual probes across L1–L40. Changes are relative to visual L40; outlines mark row maxima.

Light-MER. Figures 10 and 11 apply the same protocol to the shallower HuBERT-base and CLIP ViT-B/16 encoders. Both use L12 as the reference. The visual diagnostic reads full CREMA-D frames.

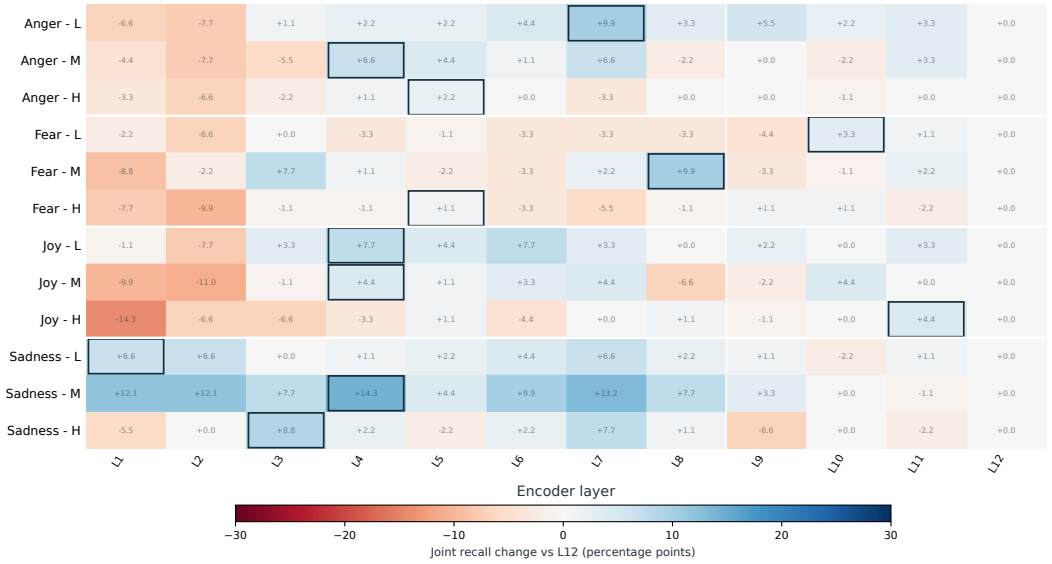


Figure 10: Light-MER HuBERT-base audio probes across L1–L12, relative to L12. Samples, actor folds, probe protocol, and color scale match the other backbones.

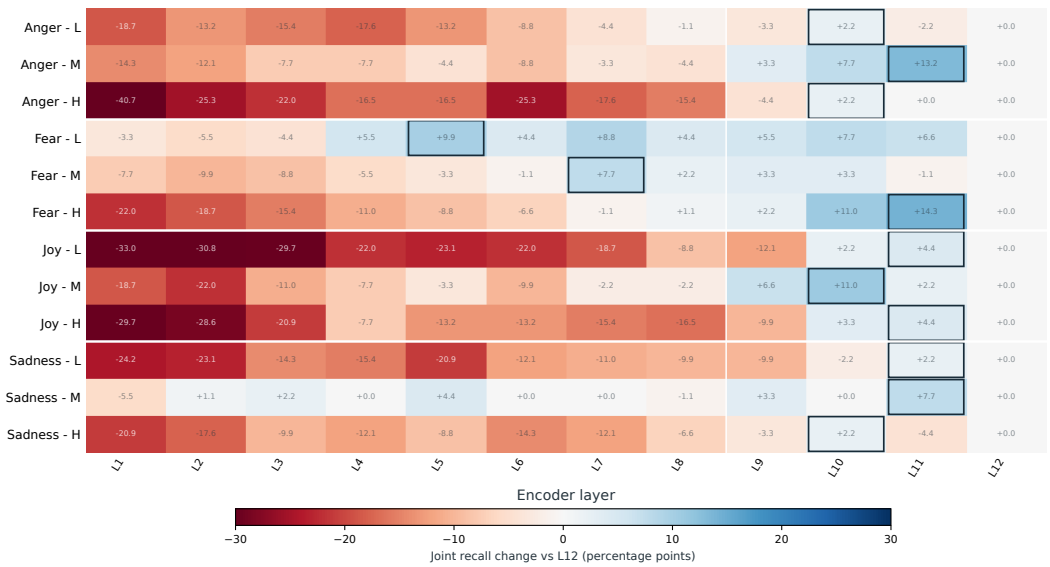
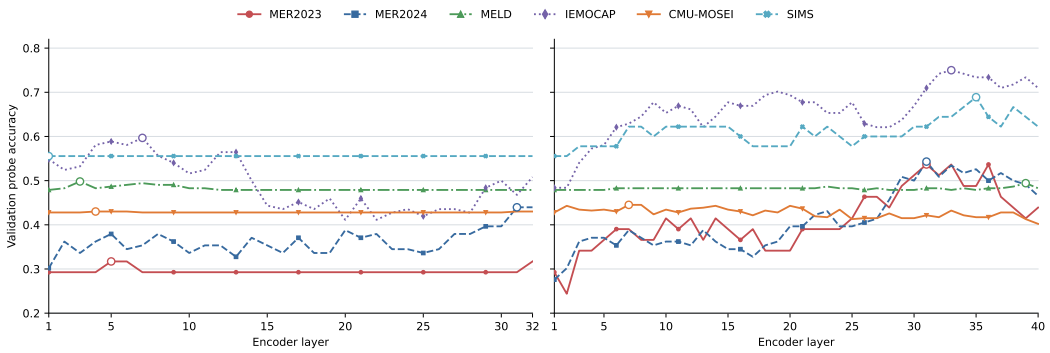


Figure 11: Light-MER CLIP ViT-B/16 visual probes across L1–L12, relative to L12. This encoder diagnostic uses full frames; the native Light-MER visual pathway uses face crops.

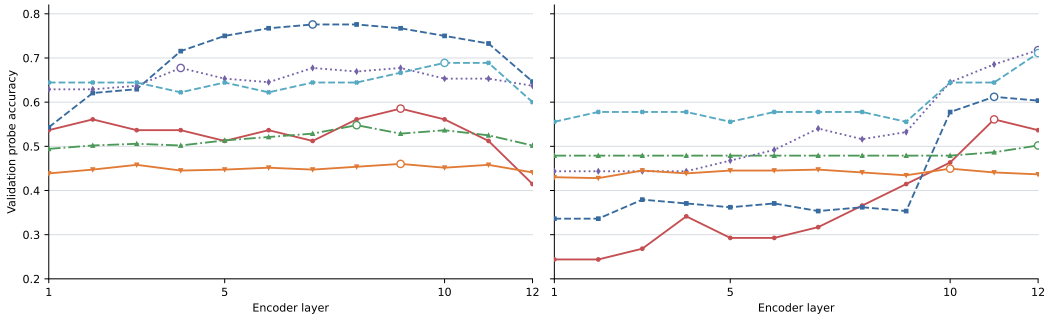
B.2 Cross-Dataset Layer-Wise Probes

We next examine how the depth profiles vary across MER2023, MER2024, MELD, IEMOCAP, CMU-MOSEI, and SIMS. Figure 12 shows audio and visual probes for Emotion-LLaMAv2 and Light-MER. Media are taken from the datasets' test/test1 partitions and internally divided approximately 9:1 for probe training and validation. Each layer has an independently trained linear head on cached frozen features; each plotted point is the highest internal validation accuracy attained during that head's training. The probe seed is 2024.

These are internal-validation depth diagnostics. The AffectGPT profiles in Figure 1(a,b) use online feature extraction with batched padding and contain five audio curves, with no CMU-MOSEI audio curve under the same protocol. The supplementary profiles therefore extend the analysis of depth preferences within each backbone. Their absolute accuracies are interpreted under these backbone-specific extraction protocols.



(a) Emotion-LLaMAv2



(b) Light-MER

Figure 12: Cross-dataset linear probes. Each row shows audio on the left and vision on the right. Emotion-LLaMAv2 uses Whisper-large-v3 L1–L32 and EVA-ViT-G L1–L40; Light-MER uses HuBERT-base and full-frame CLIP ViT-B/16 L1–L12. Hollow circles mark each curve's peak internal validation accuracy. All panels share the top legend.

C Conditional Structure of Encoder Features

Layer-wise probes measure accessible affective information. We complement these readouts with the geometry of emotion-intensity conditions, examining how their directions relate across disjoint actor groups and how their normalized responses vary with depth.

C.1 Cross-Actor Condition Directions

Figures 13–15 show cosine similarities between 12 emotion-intensity condition directions estimated from disjoint actor groups. Condition centers are fitted using 55 actors and compared with 36 held-out actors. Each figure places audio above vision and displays four depths. The common color scale is $[-1, 1]$. Diagonal entries measure cross-actor consistency of the same condition.

The matrices reveal both related conditions and modality-specific separations. Changes across the displayed depths complement the probe profiles: depth affects the organization of condition information as well as its readout utility. This heterogeneity motivates the depth-specific evidence transformations used by LayerRefit.

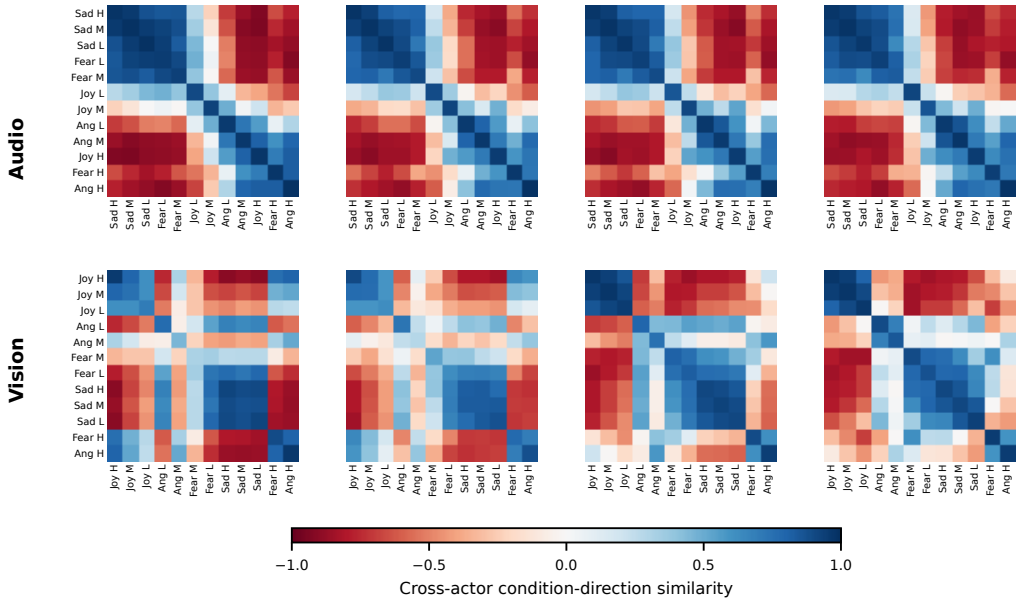


Figure 13: AffectGPT condition-direction similarities. Audio (HuBERT, top) and vision (ViT, bottom) each show L6, L12, L18, and L24 from left to right. The similarities use the 55/36 actor split.

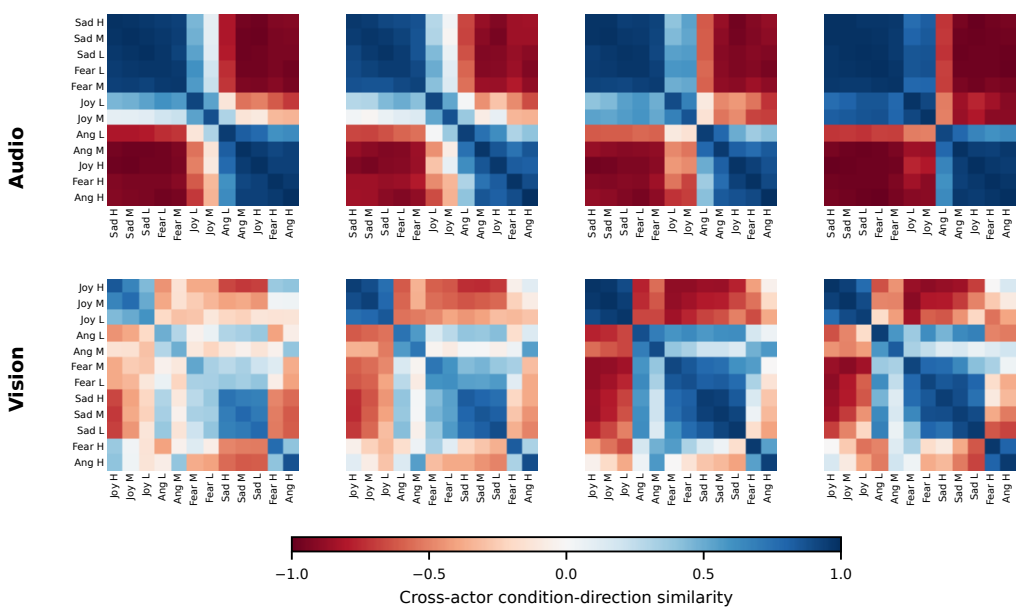


Figure 14: Emotion-LLaMAv2 condition-direction similarities. Audio (Whisper, top) shows L8, L16, L24, and L32; vision (EVA, bottom) shows L10, L20, L30, and L40. Actor groups and similarity definitions match Figure 13.

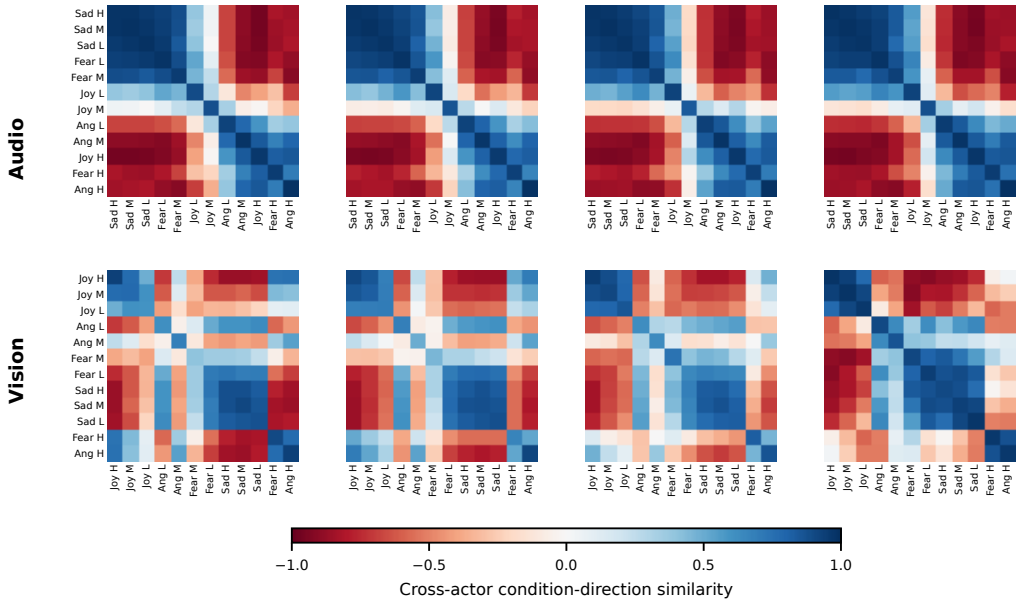


Figure 15: Light-MER condition-direction similarities. Audio (HuBERT, top) and vision (CLIP, bottom) each show L3, L6, L9, and L12. Visual features are extracted from full frames.

C.2 Conditional Response Profiles

Figure 16 summarizes modality-normalized condition responses across depth. It complements the pairwise direction matrices by displaying where each condition produces stronger responses within a modality. The six panels share the 1,092 CREMA-D clips, 91 actors, and cross-actor statistics. Response magnitude and probe accuracy provide complementary descriptions of these features.

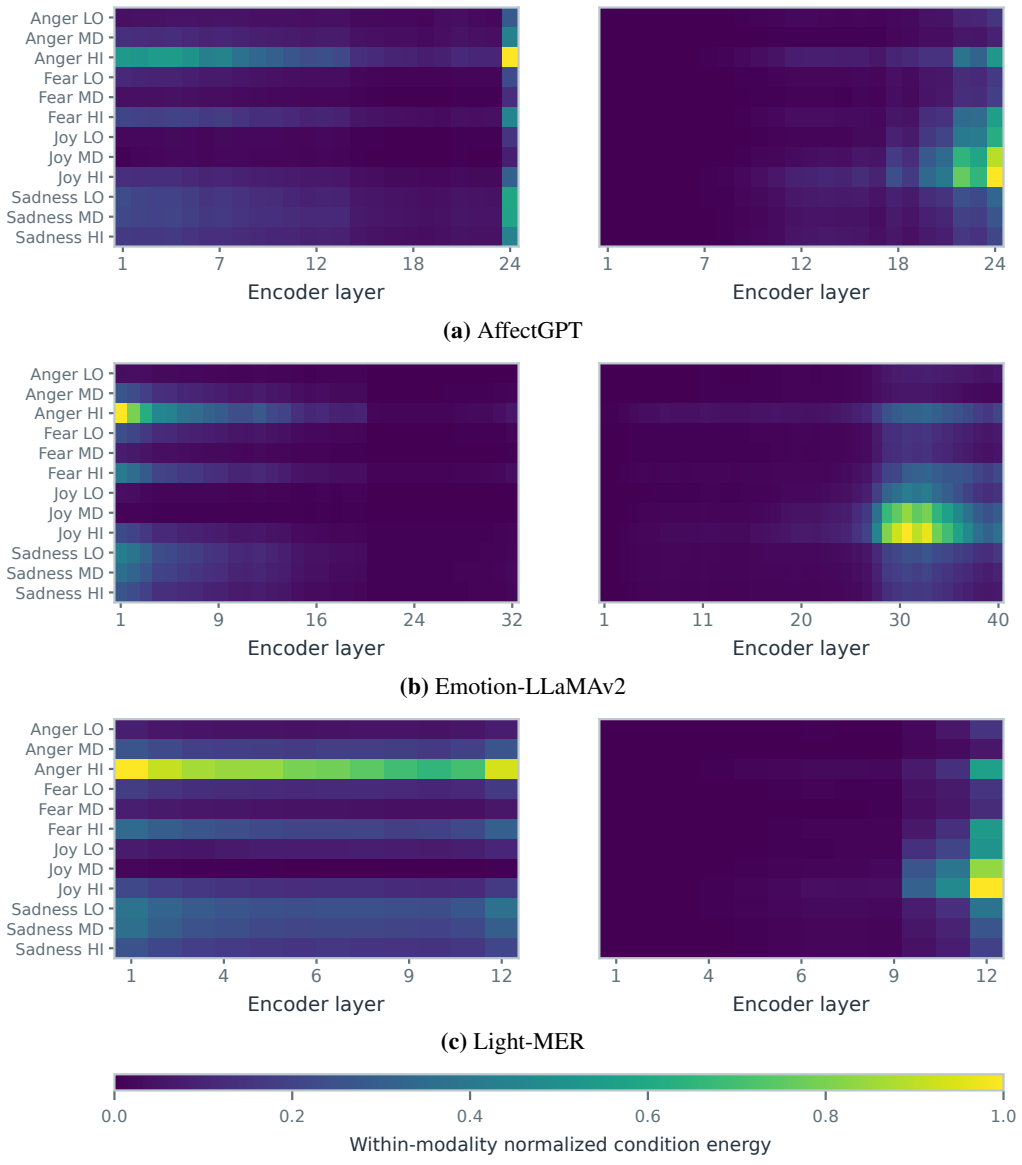


Figure 16: Condition responses across encoder depth, normalized within each modality. Each row shows audio on the left and vision on the right. AffectGPT spans L1–L24 in both modalities; Emotion-LLaMAv2 spans audio L1–L32 and vision L1–L40; Light-MER spans L1–L12 in both modalities. Light-MER visual features use full frames. The shared colorbar appears below the panels.

D Depth Access and Correction Capacity

The encoder analyses motivate access to heterogeneous depth evidence. We now examine three aspects of the refinement design: the number of selected depths, the rank needed to retain local loss-reducing directions, and the number of edit locations in the frozen model.

D.1 Number of Selected Audio and Visual Depths

We vary K_a and K_v over $\{1, 2, 3, 4, 6\}$. Figure 17 reports Macro-F1 and NLL changes relative to the unrefined $(0, 0)$ configuration. Positive ΔF1 denotes higher Macro-F1; negative ΔNLL denotes lower true-label loss. The marked configurations are $(2, 6)$ for AffectGPT, $(3, 6)$ for Emotion-LLaMAv2, and $(2, 6)$ for Light-MER, with the layer IDs in Table 6.

The grid separates the number of accessed depths from their observed utility. Larger layer sets do not yield uniformly larger Macro-F1 gains, and configurations with lower NLL need not have the highest Macro-F1. This complements the main-text ablations: access to additional evidence is useful when its contributions are organized into effective corrections.

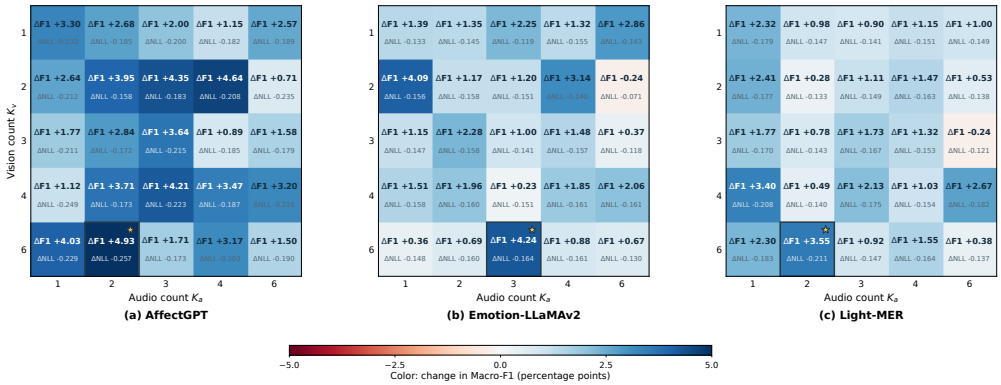


Figure 17: Changes with the number of selected depths for (a) AffectGPT, (b) Emotion-LLaMAv2, and (c) Light-MER. Each cell reports the configuration's Macro-F1 change in percentage points and NLL change relative to $(0, 0)$. The shared, zero-centered color scale encodes Macro-F1 change; blue indicates improvement and red indicates decline. Gold stars and borders mark the configurations listed in Table 6.

D.2 Local Oracle Geometry on Source Data

The oracle diagnostic examines whether local loss-reducing directions induced by true-label gradients concentrate in a compact subspace. Bases are fitted on source-training records and evaluated on disjoint held-out records, with local interventions fixed at 1% RMS amplitude. The left panels report NLL improvement retained relative to rank 64; the right panels compare held-out gradient energy captured by gradient SVD, activation PCA, and random bases. These measure task-loss utility and gradient geometry, respectively.

The oracle uses true-label gradients to characterize available local directions; the learned refinement branch generates corrections from encoder evidence. This separates the geometric motivation from the trained LayerRefit evaluations in the main text.

The Light-MER results illustrate the distinction between energy and task utility: rank 32 retains over 85% of rank-64 NLL improvement in both modalities while capturing under 28% of gradient energy. A compact space can therefore preserve substantial local task utility without capturing most gradient energy. Together with the main-text ablations, these results motivate configuring evidence access and correction capacity separately.

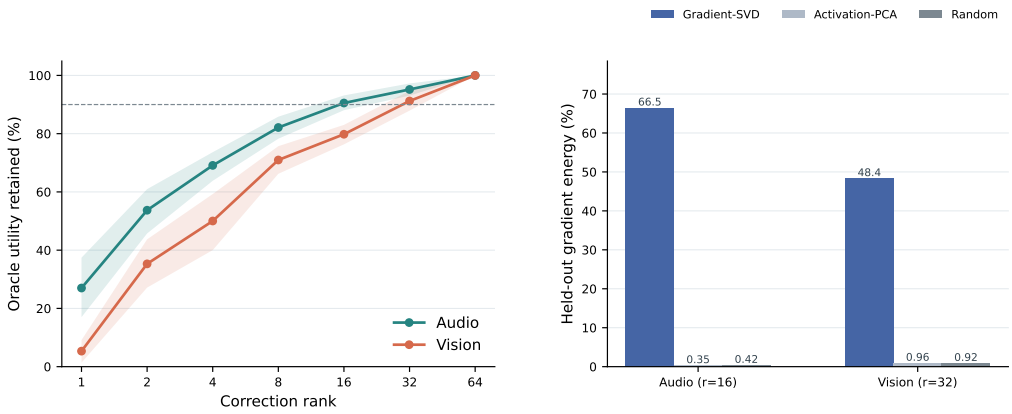


Figure 18: AffectGPT local oracle geometry on MER-Caption+ source records. At 1% RMS intervention amplitude, audio rank 16 and vision rank 32 retain 90.5% and 91.2% of rank-64 NLL improvement, respectively (left). Held-out gradient energy is shown on the right.

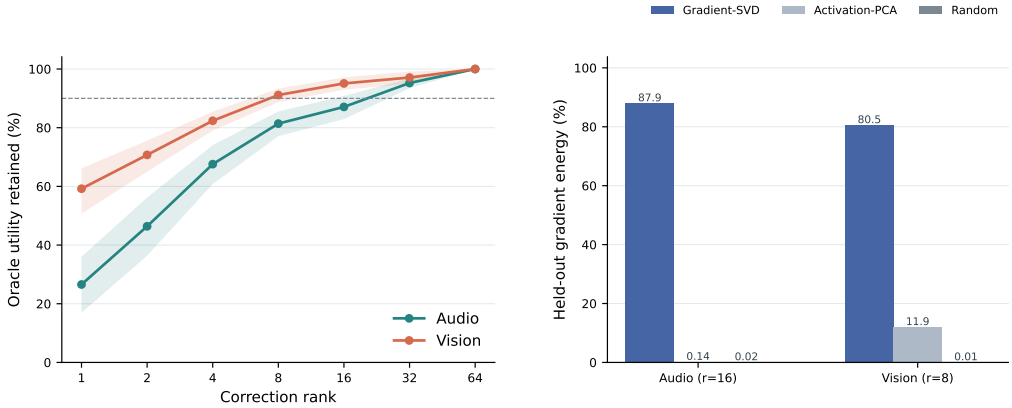


Figure 19: Emotion-LLaMAv2 source-data oracle diagnostic with one seed. Interventions of 1% RMS are applied to audio and visual features after a frozen common-basis correction. Audio rank 16 and vision rank 8 retain 87.1% and 91.1% of rank-64 NLL improvement; their retained gradient energy is 87.9% and 80.5%, respectively.

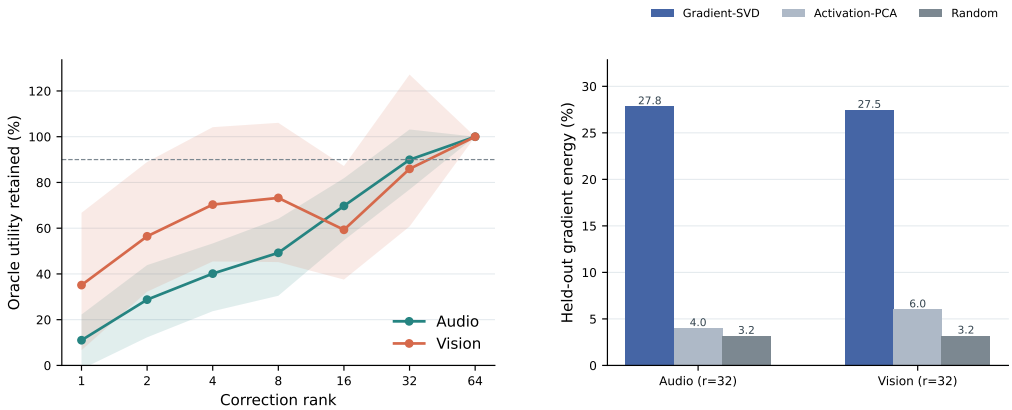


Figure 20: Light-MER source-data oracle diagnostic with one seed, using 1% RMS interventions on native audio and visual query tokens. At rank 32, audio and vision retain 89.9% and 85.9% of rank-64 NLL improvement, and 27.8% and 27.5% of gradient energy, respectively. Shading on the curves shows paired-sample bootstrap intervals.

D.3 Number of Edit Locations

LayerRefit applies one correction at the post-connector interface, before the first LLM block. Let J denote the number of edit locations: $J = 1$ is this input interface, and $J > 1$ adds internal LLM locations. Table 10 reports training and test F1 for the location sweep. Training F1 rises with J, whereas test F1 generally declines, with a small local increase from $J = 2$ to $J = 3$.

Table 10: Number of edit locations and training/test F1. $J = 1$ applies one input-interface correction; larger J adds internal LLM edit locations.

J	Edit locations	Training F1 (%)	Test F1 (%)
1	Input interface only	69.42	61.87
2	Input interface + 1 internal	71.06	61.22
3	Input interface + 2 internal	71.91	61.36
4	Input interface + 3 internal	73.12	60.17
8	Input interface + 7 internal	76.38	59.29
16	Input interface + 15 internal	78.05	58.76

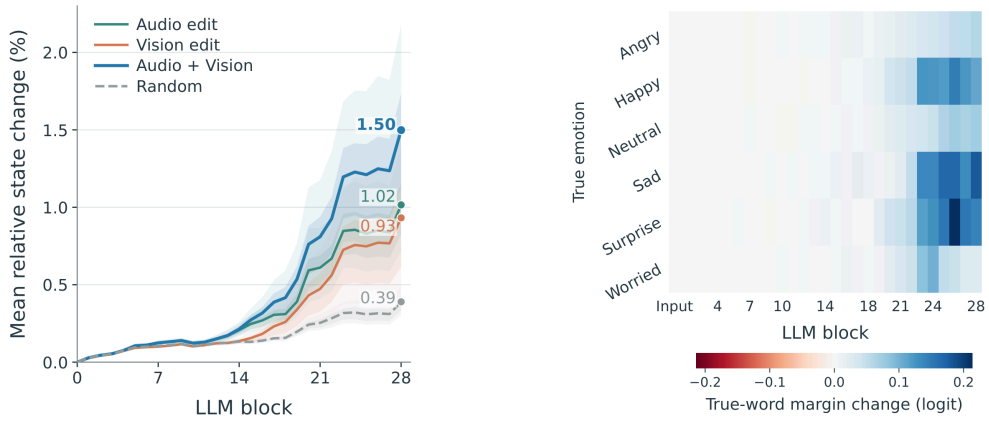
This sweep shows a tradeoff between training fit and test recognition as additional locations expand editing capacity. LayerRefit concentrates the update at the native input interface and relies on the frozen LLM for subsequent propagation. The edit-location ablations in Table 3 provide a complementary comparison.

E Propagation to the Answer Context Across Backbones

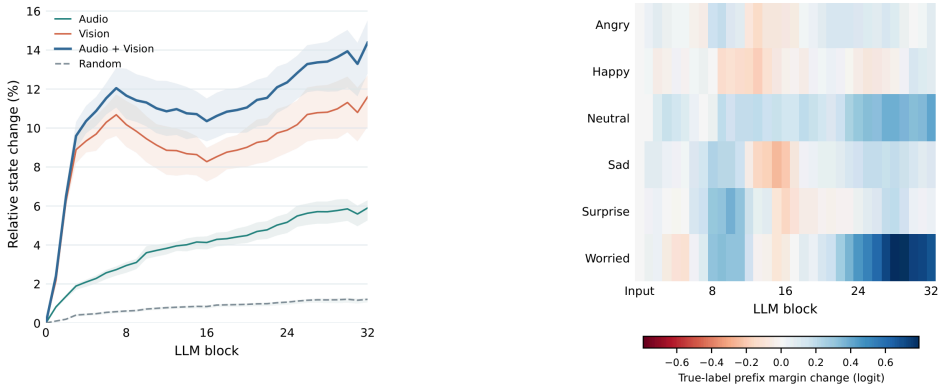
The preceding analyses concern encoder evidence and correction capacity. Figure 21 follows the resulting changes through the frozen LLM to the final context position before answer generation, extending Figure 4 to all three backbones.

For each LLM block, the left panels measure relative hidden-state change between refinement on and off, $100\|h_{\mathrm{edit}} - h_{\mathrm{off}}\|/\|h_{\mathrm{off}}\|$. Curves show sample means and shading shows the interquartile range. Audio-only, vision-only, joint, and random-direction edits are compared. The right panels use the frozen output head to read changes in class-conditioned margins across blocks: true-word margins for AffectGPT and true-label prefix margins for Emotion-LLaMAv2 and Light-MER. Emotion-LLaMAv2 and Light-MER each use six source-validation samples per class, for 36 samples per backbone. Intermediate-block margins are diagnostic readouts.

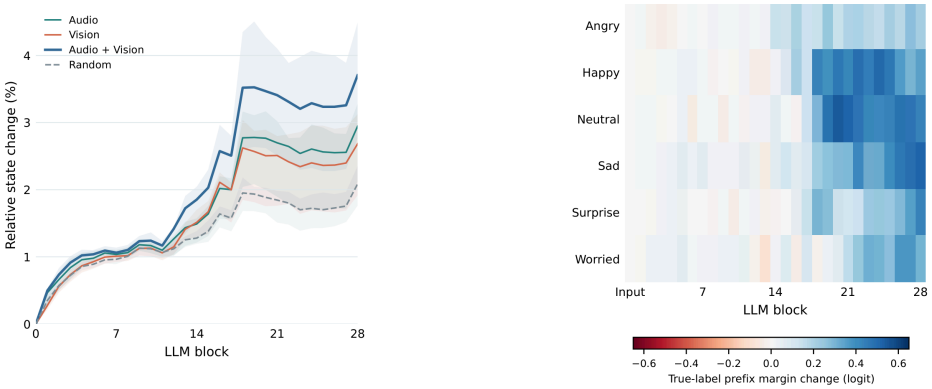
Across these profiles, joint edits generally produce larger answer-context responses than either modality alone, while class-conditioned margins vary across blocks. These diagnostics connect input corrections to answer readouts; the main-text task metrics quantify recognition gains.



(a) AffectGPT



(b) Emotion-LLaMAv2



(c) Light-MER

Figure 21: Answer-context state responses (left) and output-head margin changes (right) in three frozen backbones. Left-panel curves compare audio, vision, joint, and random-direction edits; shading gives interquartile ranges. Emotion-LLaMAv2 and Light-MER each use 36 source-validation samples. Readouts and sample sources are backbone-specific.