EMOTION INTERPRETABILITY IN MULTIMODAL MODELS: DISENTANGLING BEHAVIORAL, REPRESENTATIONAL, AND INTERVENTIONAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reported emotion circuits controlling expression and emotion neurons shared between speech and faces suggest a shared, steerable emotion representation in multimodal models. This claim bundles three kinds of evidence that benchmark accuracy cannot separate: behavioral (which input carries emotion), representational (whether text-learned emotion is recognized in faces), and interventional (whether it steers generation). We test each separately, comparing transcripts with video frames, classifying faces with emotion prototypes built from transcripts, and steering with neuron groups found without emotion labels or a simple mean-difference direction. Transcripts outperform single frames in three vision-language models, by 4.82 to 12.69 percentage points. Encouragingly, text-learned emotion carries over to faces: the prototypes beat a shuffled-label baseline in Qwen2.5-VL-7B and, after feature standardization, in Qwen3-VL-32B, and a text-derived joy direction raises the probability of labeling a face as joy by 30.07 and 36.71 points. At the same time, steering depends on the method and registers more with a classifier than a human reader. Neuron groups give no significant gain across 90 conditions in three text models, whereas the mean-difference direction raises classifier-judged emotion in three of five models, lifting surprise by 69.27 points in Qwen3-32B. A blinded rater judged all 54 sampled mean-difference continuations readable and named the target emotion in 2. We study six open-weight models, 1,411 conversational clips, and 240 acted face clips. Claims of a shared, steerable emotion representation are therefore best tested as three claims, with steering effects confirmed by human readers.

1 INTRODUCTION

Emotion labels attached to conversation can reflect words, facial expressions, prosody, and context. A transcript and a frame from the same utterance need not expose the same evidence. This creates an interpretive problem for multimodal evaluation: a difference in classification accuracy may reflect what the input shows, how the model uses it, or both. Locating the difference requires representational and interventional evidence. Heterogeneous emotion benchmarks make this distinction especially relevant because their constituent datasets differ in annotation task and modality (Hu et al., 2025; Zhang et al., 2026b; Poria et al., 2019; Lian et al., 2023).

A second question is when a predictive feature is also a steering handle. Recent work reports emotion circuits that control emotional expression more accurately than prompting- and steering-based methods (Wang et al., 2026), emotion-sensitive neurons whose interventions transfer causally between speech and faces (Zhao et al., 2026), and unsupervised neuron modules with meaningful semantics (Yu et al., 2026). Together these results raise the question whether units found through unsupervised activation structure, rather than through emotion labels or causal search, can also serve as cross-modal steering handles for emotion. We test that question directly: whether units selected by co-activation structure steer emotion, whether text-derived emotion evidence reaches vision, and whether the outcome measure itself is valid, since a classifier can assign an emotion to damaged text and a likelihood gate can accept repetition.

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.

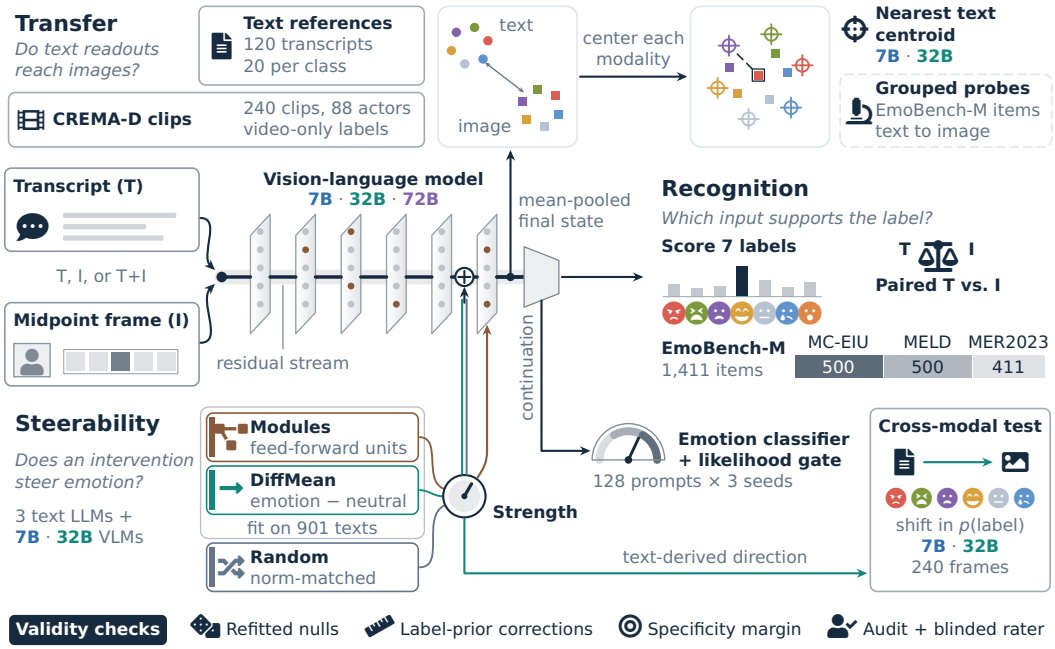

Figure 1: Study overview. Recognition scores seven labels by answer log probability from a transcript (T), a midpoint frame (I), or both (T+I) on 1,411 EmoBench-M items in three Qwen vision-language models. Cross-modal transfer centers each modality and assigns 240 CREMA-D frames to the nearest centroid of 120 reference transcripts in Qwen2.5-VL-7B and Qwen3-VL-32B; grouped probes train on text and test on images of the 1,411 items. Point clouds are schematic. Steerability is tested by amplifying co-activation modules over a strength grid in three text models and two vision-language models; the matched comparison with a mean-difference (DiffMean) or norm-matched random direction selects strengths on 32 development prompts in all five.

We therefore distinguish three kinds of evidence, each tied to one question (Figure 1). **Recognition** (behavioral): which input condition supports correct labels on matched items? **Cross-modal transfer** (representational): does an emotion readout learned from text predict visual labels without fitting to those test images? **Steerability** (interventional): does manipulating a candidate mechanism increase the intended emotion relative to an unsteered baseline, without larger off-target shifts or loss of readability? We examine these questions through complementary experiments and measure all three in Qwen2.5-VL-7B and Qwen3-VL-32B.

We show where these forms of evidence agree and where they diverge. First, on 1,411 items, a transcript advantage in three vision-language models survives two corrections for label bias, while MER2023 favors frames even after removing each model's average label preference. Second, class-level text-to-image transfer on visually annotated clips depends on how modality offsets are handled: a readout that removes only the text mean collapses onto one class for Qwen3-VL-32B and three for Qwen2.5-VL-7B, centering each modality gives six-class transfer above a shuffled-label null for Qwen2.5-VL-7B, and standardizing each modality does so in both models. Third, candidate feed-forward modules in three text models yield no significant gain in the gated target-emotion rate at any tested strength, whereas a simple mean-difference direction under the identical classifier, gate, prompts, and seeds produces large classifier-judged gains that exceed every off-target shift for joy and surprise in Qwen3-32B and Qwen3-VL-32B, smaller gains in Qwen2.5-72B, and none in Llama-3.3-70B or Qwen2.5-VL-7B; a blinded human rater finds sampled continuations from all five models readable and rarely names the target emotion. In Qwen3-VL-32B a module and two random directions also raise classifier-judged rates, which shows why the random-direction control matters. In both vision-language models a text-derived joy direction shifts image-only judgments. Finally, a metadata-masked AI-assisted review, whose readability ratings the rater matches on 11 of 12 outputs, shows that the likelihood gate accepts degenerate output.

2 Related Work

Emotion in multimodal models. Emotion benchmarks for multimodal models span many scenarios and tasks (Hu et al., 2025; Zhang et al., 2026b), and trained systems such as Emotion-LLaMA fuse audio, visual, and textual encoders (Cheng et al., 2024). Multimodal models also rely heavily on text across modalities (Wu et al., 2025a). Our single-input comparison complements these aggregate views: the transcript advantage survives calibration of zero-shot label scores (Zhao et al., 2021), and frames are more informative on one constituent source.

Shared representations and cross-modal transfer. Semantically equivalent inputs from different languages and modalities share intermediate representations, and interventions in one data type change outputs in another (Wu et al., 2025b); text-derived steering vectors improve visual task accuracy (Gan et al., 2026), and emotion-sensitive neurons transfer between speech and faces (Zhao et al., 2026). Image and text embeddings of contrastive models occupy separate regions (Liang et al., 2022), and subtracting each modality's mean closes this modality gap (Zhang et al., 2024). We show that this correction restores six-class predictions from a collapsed emotion readout and gives above-null transfer for Qwen2.5-VL-7B, and that a text-derived joy direction shifts image-only judgments. Following probe-control work (Hewitt & Liang, 2019; Belinkov, 2022), we test every transfer and probe accuracy against nulls refit on shuffled labels.

Emotion mechanisms and steering. Emotion representations in language models are localized and steerable through appraisal concepts (Tak et al., 2025), causally identified circuits control emotional expression (Wang et al., 2026), and steering visual emotional information flow improves emotion understanding in vision-language models (Zhang et al., 2026a). Iterative Decoupling (IterD) discovers semantically meaningful neuron modules by alternating neuron and sample reassignment under an activation-modularity and balance objective (Yu et al., 2026); our IterD-style variant uses bipartite modularity of the rectified activations and adds an emotional-versus-neutral contrast penalty (Appendix A). Activation addition, contrastive mean differences, and inference-time intervention are standard steering methods (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023), steering effects must be weighed against other output properties and capabilities (Li et al., 2023; Huang et al., 2026b), AxBench compares steering proposals directly with simple baselines and finds prompting strongest (Wu et al., 2025c), and high likelihood favors bland, repetitive text (Holtzman et al., 2020). We evaluate co-activation modules against a mean-difference baseline under one measurement, with a quality audit and a human check.

3 Experimental Design

3.1 Conversational recognition and visual observability

Conversational set. We use the transcript-bearing emotion subsets of EmoBench-M: 500 MC-EIU (Liu et al., 2024), 500 MELD, and 411 MER2023 items. The seven labels are anger, disgust, fear, joy, neutral, sadness, and surprise. The majority class is neutral (255/1,411; 18.07%). Text comes from the benchmark transcripts, including the benchmark's supplied language rendering. Sentiment-only items and emotion-revealing audio descriptions are excluded.

Each item is evaluated with its transcript (T), a frame nearest 50% of clip duration (I), or both (T+I). Frame selection is fixed independently of emotion predictions. Models receive the same seven-label question, with neither audio nor native video input, so I is a single-frame condition. Frames are used uncropped and unselected for expressive intensity, so they can include scene context or embedded text.

The models are Qwen2.5-VL-7B-Instruct, Qwen3-VL-32B-Instruct, and Qwen2.5-VL-72B-Instruct (Bai et al., 2025b;a), with pinned, unquantized bfloat16 weights. Classification selects the label with the highest answer log probability; every label is a single answer token, so mean and summed token scores coincide. Appendix A records prompts and runtime details.

Visually annotated comparison. To tie labels more closely to visual evidence, we select 240 CREMA-D clips (Cao et al., 2014), balanced over six emotions (40 each; no surprise) and spanning 88 actors. Selection requires a video-only human majority label with at least 60% agreement and at

least five votes. These are human labels of whole clips. The primary input is the midpoint frame; a sensitivity condition supplies frames at 25%, 50%, and 75%. CREMA-D actors read the same 12 sentences in every emotion (Cao et al., 2014), so its own text carries no emotion information; a separate set of 120 conversational transcripts, balanced at 20 per class, instead supplies the text references. This experiment therefore tests class correspondence across datasets rather than alignment of paired utterances.

Behavioral inference. We report accuracy and macro-F1. Paired correctness is compared by two-sided exact McNemar tests with Holm correction over the three primary T–I comparisons, and 10,000-draw group-bootstrap intervals resample MELD dialogues, other-source videos, or CREMA-D actors. Source breakdowns are descriptive.

Label-prior controls. Zero-shot label scoring can prefer some answer words regardless of input. We therefore rescore the saved answer scores in two ways that use no gold labels. Label-free prior removal subtracts, per model and input condition, the mean log-softmax label score across items. A log-space variant of contextual calibration (Zhao et al., 2021) subtracts the mean log-softmax scores of content-free inputs: the same question with "N/A", "[MASK]", or "..." as transcript and a flat gray, black, or white image.

3.2 Representational transfer

Text-centroid transfer. We extract the final hidden state averaged over input tokens before answer scoring. Class centroids are fitted only on the 120 text references, and each test image is assigned to the nearest centroid by cosine similarity; no image labels enter fitting. Because the offset between modality means dominates the gap in contrastive image-text embeddings (Liang et al., 2022; Zhang et al., 2024), the primary readout, specified before this analysis was run, centers each modality on its own mean; the mean of the unlabeled test images uses no labels. We also report a readout that removes only the text-reference mean and a sensitivity readout that also divides each modality by its per-dimension standard deviation. For every readout we compare accuracy against 1,999 permutations of reference labels, refitting centroids each time, bootstrap over test actors, and report the predicted-class distribution. Holm correction covers the primary readout in two models. We also compare text and image class-distance geometry using Spearman correlation and all $6! = 720$ category permutations.

Grouped probes. On the 1,411 conversational items, logistic probes use $C = 1$, five stratified group folds, and training-fold standardization. A text-to-image probe trains on training-fold text features and predicts images of held-out items using the text-fitted scaler; a second variant standardizes images with unlabeled image training-fold statistics. We examine full final-layer features for all three models and Qwen3-VL-32B group signatures, which average absolute feed-forward down-projection inputs across tokens within each of six co-activation groups (G0–G5) from a separate historical partition (64 layers by six groups), in six-dimensional layer-averaged and 384-dimensional concatenated layouts. The null shuffles labels within groups and repeats the entire split–scale–fit procedure for 1,999 permutations, so the smallest raw p is 0.0005. Shuffling can relabel 401 items in 117 mixed-label groups, so this conditional null tests within-group label assignment. We report Holm correction over the 15 original and all 20 tests, and binomial tests against the majority rate.

3.3 Interventions and generation quality

Candidate modules. For Qwen3-32B, Llama-3.3-70B-Instruct, and Qwen2.5-72B-Instruct (Yang et al., 2025; Grattafiori et al., 2024; Qwen Team et al., 2024), we retain five modules per model from the existing discovery partitions of 901 texts. A module targets the most common non-neutral label among its texts, so targets repeat. For the same-model analysis we rerun discovery with identical settings in the language towers of Qwen2.5-VL-7B and Qwen3-VL-32B on the same 901 texts, which yields three anger and two sadness modules, and two anger, two joy, and one sadness module, respectively. For selected units u, intervention changes the input to the feed-forward down projection as

$$h'_u = h_u + \alpha\sigma_u, \tag{1}$$

where σ_u is the discovery-token activation standard deviation. The historical grid, fixed in the analysis protocol before the 128-prompt runs, uses $\alpha = 4$ as primary and $\alpha \in \{2, 8\}$ as sensitivity settings; we extend it to $\alpha \in \{0.25, 0.5, 1\}$ for all three text models. Each condition has 128 neutral continuation prompts and three paired seeds, with an unsteered baseline for the same prompt and seed. Generation uses temperature 0.8, top-$p = 0.95$, and at most 64 new tokens. The 384 outputs per condition yield 128 prompt-level inferential units after averaging seed differences.

Mean-difference baseline and strength selection. For target e and decoder layer ℓ, the DiffMean direction $v_{e,\ell}$ (Wu et al., 2025c) is the unit-norm difference between the mean residual stream leaving layer ℓ for discovery texts labeled e and for neutral discovery texts; during generation we add $c\,\bar{n}_\ell v_{e,\ell}$ at every position, where $\bar{n}_\ell$ is the mean residual token norm. A control adds a norm-matched random Gaussian direction. Layer, strength, and module α are selected only on 32 separate development prompts by a fixed gate-constrained rule (Appendix A), then evaluated once on the 128 prompts and three seeds in all five models, with Holm correction over the 17 conditions of each model.

Outcome and specificity. An external seven-class English emotion classifier (Hartmann, 2022) judges the saved continuations. A separate DistilGPT2 language model (Sanh et al., 2019) supplies a quality proxy: mean negative log likelihood must not exceed the 95th percentile of 32 separate unsteered calibration generations for that model. We call this a *gate pass* and test its validity as a fluency measure in Section 6.4. For target e and saved continuation x, the gated steering success rate is

$$\mathrm{SSR}_e = \mathbb{E}[\mathbf{1}\{\widehat{e}(x) = e\}\mathbf{1}\{g(x) = 1\}], \qquad \Delta\mathrm{SSR}_e = \mathrm{SSR}_e^{\mathrm{steered}} - \mathrm{SSR}_e^{\mathrm{base}}. \tag{2}$$

Two-sided sign-flip tests use 19,999 prompt-level draws; confidence intervals bootstrap prompts. To detect off-target effects, we report

$$M_e = \Delta f_e - \max_{j \neq e} \Delta f_j, \tag{3}$$

where f_j is the classifier's ungated argmax-label frequency, with a prompt-bootstrap interval. A steered gate-pass rate below 10% forces $\mathrm{SSR}_e^{\mathrm{steered}}$ toward zero regardless of emotion content, so we label such decreases gate-collapse floors. From the real baseline records we also simulate the smallest ΔSSR that the design detects with 80% power.

Cross-modal intervention test. In Qwen2.5-VL-7B and Qwen3-VL-32B, we add the text-derived direction for each non-neutral CREMA-D emotion, or a norm-matched random direction, while the model classifies the 240 midpoint frames without text. The outcome is the change in the probability of the target label, tested by actor-level sign flips against the random direction (Appendix A).

Supplementary checks. A post-hoc screen flags repetition and nonlinguistic structure in the 17,280 historical module-intervention outputs ($\alpha \in \{2, 4, 8\}$) and the 34,560 matched-evaluation outputs, a metadata-masked AI-assisted review rates 135 baseline/intervention pairs for readability and continuation compliance (Appendix E), and a blinded independent human rater checks that review (Appendix G). We re-extract the three text models' discovery activations on the pinned revisions and score each stored partition under its original objective; a zero-ablation pilot is in Appendix F.

4 Recognition Depends on the Input Source

All three models show an aggregate transcript advantage (Appendix B, Table 3). Qwen2.5-VL-7B is correct on 444 transcripts and 376 frames, a 4.82 percentage-point gap. Qwen3-VL-32B scores 555 versus 376 (12.69 points), and Qwen2.5-VL-72B scores 541 versus 363 (12.62 points). All three paired tests survive correction, and their group-bootstrap difference intervals exclude zero.

The aggregate hides a consistent source reversal (Figure 4; Table 4). MELD favors transcripts by 17.00, 27.00, and 26.00 points for the 7B, 32B, and 72B models, respectively, and MC-EIU by 2.20, 12.00, and 11.40 points. MER2023 instead favors frames by 6.81, 3.89, and 2.19 points. Combining transcript and frame gives 33.31%, 37.92%, and 37.92% accuracy, with no significant paired T–(T+I) difference (uncorrected $p = 0.173$, 0.349, and 0.786). The frame's contribution depends on the source: it raises accuracy over the transcript alone on MER2023 by 10.46, 8.27, and 12.41 points and

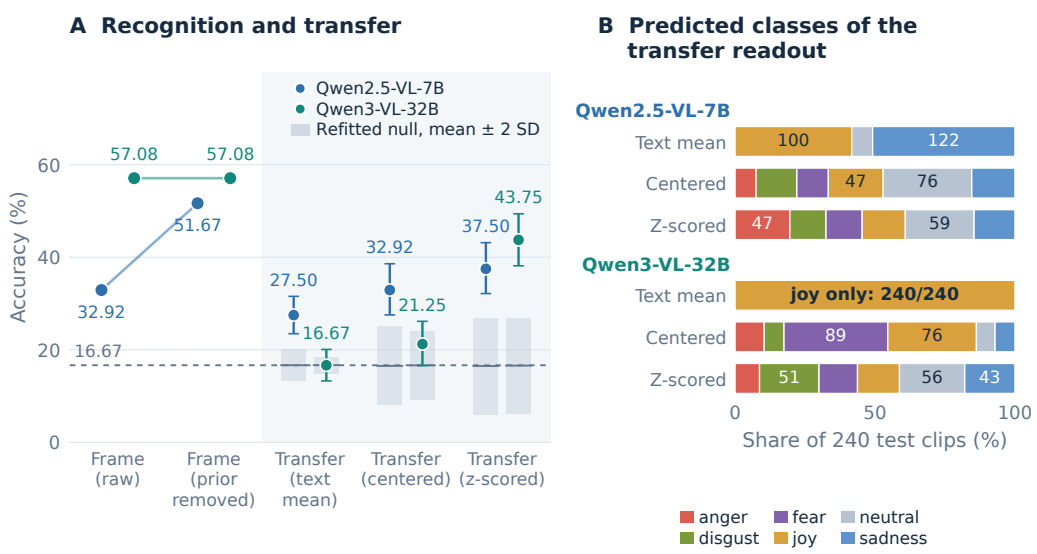

Figure 2: Recognition and text-to-image transfer on 240 CREMA-D clips from 88 actors. (A) Midpoint-frame recognition under raw scoring and after label-free prior removal, and transfer accuracy under three readouts, with actor-bootstrap intervals for transfer; bands mark the shuffled-label null mean ± 2 SD (1,999 refitted permutations). The dashed line is the balanced 16.67% reference. (B) Predicted-class distributions of each transfer readout. The text-mean-only readout assigns all 240 images to joy for Qwen3-VL-32B and uses three classes for Qwen2.5-VL-7B.

lowers it on MELD by 5.80, 10.40, and 9.60 points (Table 4), so opposite source effects cancel when pooled.

Raw frame predictions concentrate on one answer (Qwen2.5-VL-7B and Qwen2.5-VL-72B predict neutral for 907 and 824 of 1,411 frames, against 255 neutral items; Table 6), but the transcript advantage survives both corrections (Table 5): the T–I gaps are 7.44, 9.64, and 10.84 points after label-free prior removal and 6.31, 13.96, and 12.26 points after content-free calibration (all Holm $p < 0.001$). The MER2023 frame advantage persists after label-free prior removal in all three models (9.00, 7.79, and 6.57 points; Appendix B), so which input is more informative depends on the source dataset; transcript quality and translation, label composition, scene context, and visual observability are candidate explanations.

5 Cross-Modal Transfer Depends on the Readout

On visually annotated clips, raw midpoint-frame accuracy is 57.08% for Qwen3-VL-32B and 32.92% for Qwen2.5-VL-7B (Figure 2A; Table 10). Much of this gap is a label prior: Qwen2.5-VL-7B predicts neutral for 170 of 240 frames, and label-free prior removal raises its accuracy to 51.67% while leaving Qwen3-VL-32B at 57.08%. Reference-text accuracy is 41.67% for Qwen3-VL-32B and 35.83% for Qwen2.5-VL-7B, so text centroids are noisy targets (three-frame results: Appendix B).

The predicted-class distributions show how the readouts differ (Figure 2B). With only the text-reference mean removed, Qwen3-VL-32B assigns every image to joy, so its 16.67% accuracy equals that class's share, and Qwen2.5-VL-7B predicts only joy, sadness, and neutral (27.50%). Under the primary modality-centered readout, both models use all six classes. Text centroids then classify 79/240 images correctly for Qwen2.5-VL-7B (32.92%; actor interval [27.54, 38.60]%; Holm $p = 0.004$) and 51/240 for Qwen3-VL-32B (21.25%; [16.60, 26.16]%; Holm $p = 0.116$). Standardizing (z-scoring) each modality, a sensitivity readout, gives 37.50% and 43.75% (raw $p = 0.0015$ and 0.0005 against refitted nulls with means of 16.50% and 16.59%). Under the primary readout, class-level transfer is therefore supported for Qwen2.5-VL-7B. Qwen3-VL-32B exceeds its null under the standardized readout, and its chance-level text-mean accuracy reflects a one-class readout. Text and image class-

distance geometries correlate positively but not significantly (centered Spearman $\rho = 0.454$ and 0.550; raw $p = 0.143$ and 0.078).

Item-level probes decode transcripts, not frames (Appendix C; Figure 5): text probes exceed their shuffled-label nulls in all five feature sets (Holm $p = 0.010$ over 20 tests), whereas frame probes and all text-to-image probes do not; the strongest transfer, 16.51% for full Qwen3-VL-32B features (Holm $p = 0.068$), is below the 18.07% majority rate. Fixed linear probes thus find no significant item-level transfer; nonlinear readouts and other layers may differ, as probed content can localize to model-specific layers (Huang et al., 2026a).

6 Steerability Depends on the Intervention

6.1 Text-model modules produce no significant positive gain at any tested strength

Across 90 module conditions (15 modules in three text models at six strengths), no positive $\Delta\mathrm{SSR}$ survives correction within its strength family or over all 90 tests, where the smallest adjusted p for a positive estimate is 0.180 within a strength family and 0.643 over all 90 tests (Appendix D). Lower strengths preserve the gate except for Llama-3.3-70B module 1 (a 0% gate-pass rate at every $\alpha \leq 1$). The largest gain at $\alpha \leq 1$ is 3.65 points (Qwen2.5-72B module 4, $\alpha = 1$; raw $p = 0.145$; Table 15), and Llama-3.3-70B's largest is 2.08 points. No specificity interval lies above zero. Of 37 decreases at $\alpha \in \{2, 4, 8\}$, 16 are gate-collapse floors with a steered gate-pass rate below 10% (Table 16). All three re-extracted partitions reproduce their stored objectives on the pinned weights (relative modularity difference below 2×10^{-7}; module-contrast correlation 1.00), confirming that the pinned weights reproduce the activations behind the tested modules. The smallest gain the design detects with 80% power is 4–10 points, and 5–15 points at the strictest Holm level of the 15-test $\alpha = 4$ family (Table 17).

6.2 A mean-difference direction steers under the same judge

Under the identical prompts, seeds, judge, and gate, development-selected DiffMean directions raise gated target rates substantially (Figure 3A; Tables 18 and 19). In Qwen3-32B (layer 48, $c = 0.4$), four of six targets increase after correction: surprise by 69.27 points (prompt interval [64.32, 74.22]), joy by 55.73 ([49.74, 61.46]), disgust by 17.19, and anger by 10.68 (each Holm $p = 8.5 \times 10^{-4}$). For joy and surprise the specificity margin is also positive (53.39 [46.09, 60.68] and 76.56 [70.05, 81.26] points), so the target shift exceeds every off-target shift. In Qwen2.5-72B (layer 60, $c = 0.4$), surprise rises by 16.41, joy by 13.28, and anger by 4.69 points (Holm $p = 8.5 \times 10^{-4}$, 8.5×10^{-4}, and 0.032), with specificity intervals that include zero. In Qwen3-VL-32B (layer 48, $c = 0.4$), joy rises by 57.81 and surprise by 66.41 points with specificity margins of 68.75 [61.72, 75.26] and 75.78 [69.79, 80.47], and anger, disgust, and fear rise by 9.11, 12.24, and 8.85 points (Holm $p \leq 0.006$). In Llama-3.3-70B (layer 40, $c = 0.1$), no condition survives correction; the largest gain, joy (4.95 points), has Holm $p = 0.317$. The same prompts, judge, and gate therefore register large classifier-judged gains when an intervention produces them, so the module null is not explained by an insensitive measurement pipeline. In the other four models, neither norm-matched random directions (largest 6.25 points, Holm $p = 0.164$) nor the development-selected modules (largest 3.65 points, Holm $p = 1.000$) produce a positive gain that survives correction, and two Qwen3-32B modules decrease their targets. In Qwen3-VL-32B, random directions raise anger by 12.24 and sadness by 26.04 points and module 1 raises joy by 38.28 points (each Holm $p = 8.5 \times 10^{-4}$), with module 1 passing the gate on 100% of outputs against an 82.03% baseline, largely through nonlinguistic output (Section 6.4); the DiffMean joy and surprise gains have by far the largest specificity margins in this model, although module 1 (10.94 [1.30, 20.31]) and the random sadness and anger directions also have margins above zero. Over two sessions, a blinded human rater judged all 54 sampled DiffMean continuations readable but named the target emotion in 2, described 3 of the 4 sampled Qwen3-32B and Qwen3-VL-32B joy continuations as excitement, and rated all three sampled Qwen3-VL-32B random-direction outputs unreadable (Appendix G); these shifts are therefore classifier-judged and subtle to a human reader.

In Qwen2.5-VL-7B, neither selected method raises a gated target rate after correction; the largest DiffMean gain is joy (3.91 points, raw $p = 0.0995$; Appendix A explains its selection).

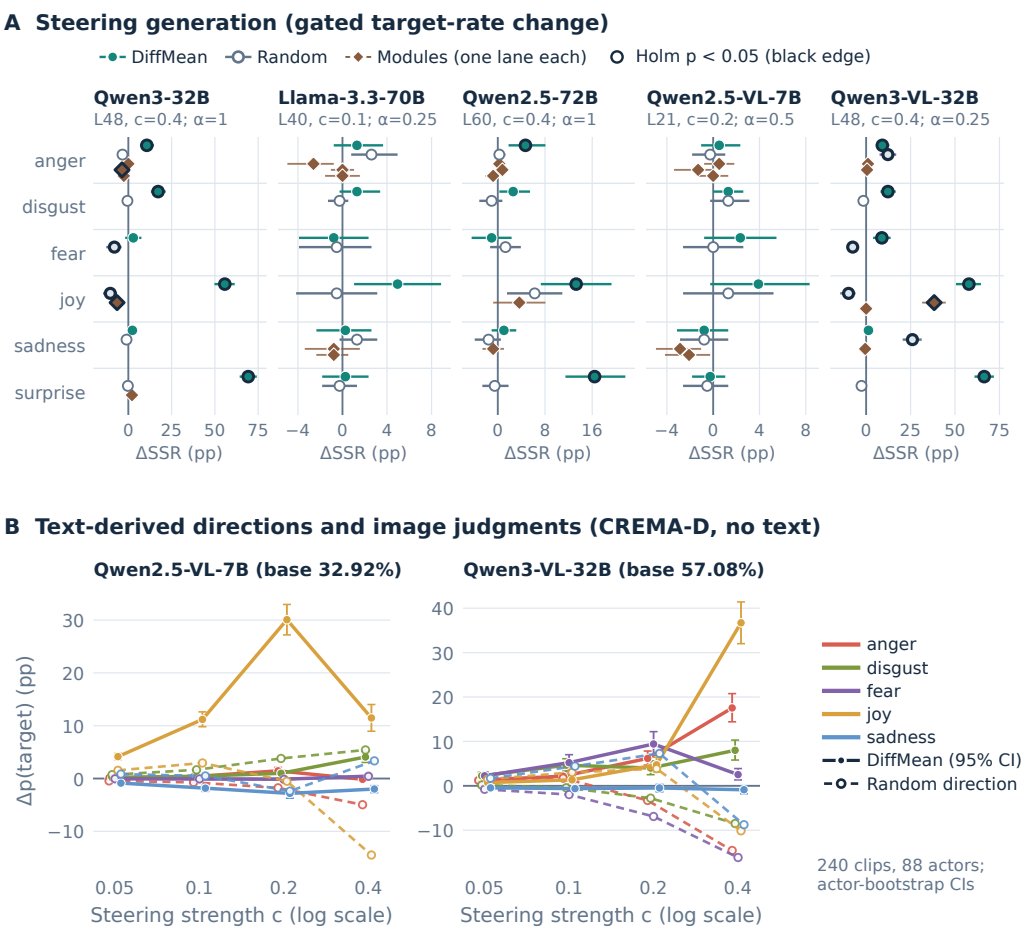

Figure 3: Steering under matched evaluation in five models. (A) Gated target-rate change (ΔSSR, percentage points) on 128 prompts × 3 seeds for the DiffMean direction, a norm-matched random direction, and each module, all at strengths selected on development prompts, with prompt-bootstrap 95% intervals; marked points pass Holm correction over the 17 conditions of each model. (B) Change in the probability of the target label when a text-derived direction (solid) or random direction (dashed) is added while each vision-language model classifies 240 CREMA-D frames without text; intervals resample 88 actors. Outcomes are automated measurements; Appendix G adds a human check of panel A continuations from all five models.

6.3 Text-derived directions shift image judgments

In the same Qwen2.5-VL-7B, adding the text-derived joy direction while the model classifies CREMA-D frames raises the probability of joy by 30.07 points at $c = 0.2$ (actor interval [27.19, 32.96]) and the share of joy predictions by 35.00 points (Figure 3B; Table 20). This shift exceeds every off-target shift (margin 30.10 points) and the random direction's effect by 30.59 points (Holm $p = 0.001$). Qwen2.5-VL-7B's own modules shift predictions nonselectively (Table 21): only sadness module 3 at $\alpha \leq 1$ has a specificity interval above zero (0.88 and 1.65 points), and module 1 lowers accuracy from 32.92% to between 16.67% and 19.17%. In Qwen3-VL-32B at the $c = 0.4$ used for generation, the joy direction raises the probability of joy by 36.71 points [32.01, 41.41] (margin 32.44; 46.84 above random) and the anger direction raises anger by 17.54 points (margin 15.82; 32.10 above random), both with Holm $p = 0.001$; the fear direction gives a smaller specific gain at $c = 0.2$ (9.42 points; margin 6.19), the disgust direction gives a small one at $c = 0.1$ (4.65 points; margin 1.93), and the sadness direction does not raise sadness. Its modules 1 (joy) and 2 (anger, $\alpha = 1$) raise their target probabilities by moving 86.67 and 70.83 points of frames to their target labels, and accuracy falls from 57.08% to 16.67% and 25.83%. Text-derived directions therefore carry emotion-specific

information into visual judgments for joy in both models and for anger, fear, and, weakly, disgust in Qwen3-VL-32B, whereas the models' own modules act nonselectively or collapse predictions onto one label.

6.4 Gate passes can conceal severe output damage

Among 17,280 historical module-intervention outputs, 9,502 pass the likelihood gate, and post-hoc screening flags 3,318 of these (34.92%); no unsteered output triggers the screen. In the AI-assisted review (Appendix E, Table 24), whose readability ratings a blinded rater matched on 11 of 12 outputs (Cohen's $\kappa = 0.875$), all 135 baselines but 23/135 steered outputs are mostly readable, and 53 of the 75 gate-passing steered outputs (70.67%) fall below mostly readable. On the 34,560 matched-evaluation outputs, the screen flags no baseline and no Llama-3.3-70B output, but 46 of the 273 gate-passing target outputs behind the Qwen3-32B surprise gain, 72 of 265 behind the Qwen3-VL-32B surprise gain, and 47 of 56 behind its random-anger gain (Table 23); it misses isolated-character strings such as Qwen3-VL-32B's module-1 outputs. A gate pass is therefore not evidence of fluent emotion steering.

7 Discussion and Conclusion

The experiments separate three kinds of evidence often bundled into a claim about shared emotion mechanisms. A transcript advantage describes prediction from particular inputs; it survives label-prior calibration, and its reversal on MER2023 shows the importance of source composition. Cross-modal transfer depends on how modality offsets are handled: centering gives above-null transfer for Qwen2.5-VL-7B, standardization gives it in both models, and the apparent contrast under the text-mean readout reflects a one-class collapse. Steerability depends on the intervention: text-model modules yield no significant generation gain at any tested strength, whereas a mean-difference direction produces classifier-judged shifts, large and with positive specificity margins for joy and surprise in Qwen3-32B and Qwen3-VL-32B, smaller in Qwen2.5-72B, and absent in Llama-3.3-70B and Qwen2.5-VL-7B; text-derived joy directions also change image judgments in both vision-language models.

The practical implications are to test readouts for degenerate predictions, to calibrate zero-shot label scores, and to compare candidate mechanisms with a working steering baseline under the same outcome measure. Encouragingly, a simple mean difference moved classifier-judged emotion in three of five models, while modules did so only in Qwen3-VL-32B, via gate-passing nonlinguistic output; yet a blinded rater rarely named the target emotion in readable mean-difference continuations. Because the likelihood gate accepts repetition and malformed output, evidence of emotion steering should rest on validated quality assessment, not likelihood alone. The same three-part design extends directly to other model families, nonlinear and layer-wise readouts, and panels of human raters.

Limitations

The vision-language models span 7B to 72B in the Qwen family; the tests extend directly to other families. Single frames isolate visual evidence from motion and sound, CREMA-D clip labels need not be visible in every frame, and benchmark exposure in pretraining is unknown. Group-aware intervals cover known speakers and dialogues; itemwise p values assume independence beyond them.

Linear readouts of pooled representations and fixed partitions keep results interpretable against refitted nulls; nonlinear, layer-wise readouts extend them. Centroid transfer uses 20 references per class, the within-group probe null relabels 28.42% of items, and standardization is a sensitivity analysis. The historical Qwen3-VL-32B partition, unlike the text-model ones, was not re-verified (Appendix A).

Prompting, the strongest steering method in AxBench (Wu et al., 2025c), would complement our activation interventions. The judge, an English classifier trained partly on MELD, is not calibrated for corrupted or non-English text, so a screen, a metadata-masked AI review, and a blinded rater (11 of 12 ratings matched) check the text it scores; more raters would extend this check. The cross-modal test measures label probabilities, and the ablation null is a five-draw pilot. The findings concern the tested models; they neither rule out module steering nor validate emotion circuits.

Ethics Statement

We use existing research datasets and their supplied annotations, redistribute no benchmark media, and report aggregate outcomes rather than identifiable face examples. The single-rater check (Appendix G) collected judgments of short model-generated texts from one rater identified only by a code. The rater, a volunteer, gave informed consent and was not paid; because the task collected no personal information and involved no deception, it did not undergo institutional ethics review. The booklet told the rater in advance that some continuations are badly damaged model outputs. No personal information about the rater is reported, and the item-level answers are withheld apart from the comments quoted in Appendix G. An emotion label represents an annotation of an utterance or performance, not a verified private mental state. The study does not infer diagnoses, identify people, or support deployment for high-stakes assessment. Directions that shift the emotional tone of generated text could be misused to produce manipulative content; we release code and summary records but no steering vectors or steered generations. Any redistribution must retain the source datasets' licenses and access restrictions.

Reproducibility Statement

The supplementary material contains the experiment and analysis code, the protocol that fixed the settings of the added transfer, probe, label-prior, steering, and cross-modal experiments before they ran, the summary records behind each added table and figure, and the rating-booklet template with the scripts that build and score it and the aggregated rating summary. Appendix A specifies checkpoints, prompts, statistical families, and reproduction details. The included CPU-only scripts rebuild every added table and figure from the included summary records, and the remaining tables and figures, except the single-rater table, from the item-level records. Item-level predictions, feature arrays, and AI-assisted review ratings are retained with source hashes for release with the code; raw generations and the rater's item-level answers, apart from the comments quoted in Appendix G, are withheld, and benchmark media remain available from their original sources.

AI Use Statement

In this work, we used AI assistant tools for grammar checking and experiment-code debugging. We also used advanced generative AI tools to get feedback on improving the human annotation interface design. We have not used generative AI tools to create data or to assign emotion labels: all stimuli and emotion labels come from existing datasets, and the emotion judge is an external classifier. As part of the analysis protocol, an LLM served as a judge in the AI-assisted readability review of sampled outputs (Appendix E), and a blinded human rater checked its ratings (Appendix G). We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts presented in this manuscript.

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. URL https://aclanthology.org/P18-1208/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76 (6):066102, 2007. doi: 10.1103/PhysRevE.76.066102.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

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.

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 (NeurIPS)*, volume 37, pp. 110805–110853, 2024. doi: 10.52202/079017-3518. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7f43ada17acc234f568dc66da527418-Abstract-Conference.html.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40056–40087, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1861. URL https://aclanthology.org/2026.acl-long.1861/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jochen Hartmann. Emotion English DistilRoBERTa-base, 2022. URL https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/. Model card.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. EmoBench-M: Benchmarking emotional intelligence for multimodal large language models, 2025. URL https://arxiv.org/abs/2502.04424.

Fan Huang, Haewoon Kwak, and Jisun An. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pp. 294–297, Austin, TX, USA, 2023. ACM. doi: 10.1145/3543873.3587368. URL https://dl.acm.org/doi/10.1145/3543873.3587368.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. ChatGPT rates natural language explanation quality like humans: But on which scales? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3111–3132, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.277/.

Fan Huang, Haewoon Kwak, and Jisun An. Understanding moral reasoning trajectories in large language models: Toward probing-based explainability, 2026a. URL https://arxiv.org/abs/2603.16017.

Fan Huang, Songheng Zhang, Haewoon Kwak, and Jisun An. CogBias: Measuring and mitigating cognitive bias in large language models, 2026b. URL https://arxiv.org/abs/2604.01366.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

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. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614. ACM, 2023. doi: 10.1145/3581783.3612836.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W. Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset, 2024. URL https://arxiv.org/abs/2407.02751.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258. ACM, 2022. doi: 10.1145/3536221.3556630.

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, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108.

Lingyun Song, Siyu Chen, Ziyang Meng, Mingxuan Sun, and Xuequn Shang. FMSA-SC: A fine-grained multimodal sentiment analysis dataset based on stock comment videos. *IEEE Transactions on Multimedia*, 26:7294–7306, 2024. doi: 10.1109/TMM.2024.3363641.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.679. URL https://aclanthology.org/2025.findings-acl.679/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do LLMs "feel"? Emotion circuits discovery and control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.11328.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models, 2025a. URL https://arxiv.org/abs/2508.10552.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://arxiv.org/abs/2411.04986.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025c. URL https://proceedings.mlr.press/v267/wu25a.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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. URL https://aclanthology.org/2020.acl-main.343/.

Yanke Yu, Jin Li, Ying Sun, Ping Li, Zhefeng Wang, and Yi Zheng. Discovering decoupled functional modules in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34503–34511, 2026. doi: 10.1609/aaai.v40i41.40749. URL https://ojs.aaai.org/index.php/AAAI/article/view/40749.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. doi: 10.1109/MIS.2016.94.

Chengsheng Zhang, Chenghao Sun, Zhining Xie, and Xinmei Tian. Interpreting and enhancing emotional circuits in large vision-language models via cross-modal information flow. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2605.21980.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://arxiv.org/abs/2508.09210.

Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations (ICLR)*, pp. 51711–51736, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e26f31de8b13ec569bf507e6ae2cd952-Abstract-Conference.html.

Xiutian Zhao, Luqi Sun, Björn Schuller, and Berrak Sisman. Emotion across speech and faces: Shared affective mechanisms in multimodal foundation models, 2026. URL https://arxiv.org/abs/2608.17102.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

A Implementation and Statistical Scope

A.1 Inputs, checkpoints, and prompts

Table 1 records the pinned data and model revisions. Frames use the closest decoded timestamp to the specified fraction, saved as JPEG at quality 95. The vision-language model (VLM) processor limits images to 401,408 pixels. The three-image sensitivity is supplied through the image interface, not through a native video processor. The fixed classification question is:

> What is the single most likely emotion being expressed? Answer with exactly one word from this list: anger, disgust, fear, joy, neutral, sadness, surprise.

For CREMA-D and its category references, surprise is omitted. The text condition prepends the benchmark transcript, the image condition supplies the frame, and the combined condition supplies both. We score each candidate continuation with the model's chat template; mean and summed log scores are saved. Full features are mean final hidden states over the entire input sequence, including instruction and visual tokens when present, before the candidate answer. Consequently, transfer can reflect distributed class information and prompt structure; it is not a probe of only face tokens or of a known emotion subspace.

Table 1: Data revisions and resolved inference checkpoints. The historical Qwen3-32B and Qwen3-VL-32B partitions do not record their discovery-time revision; the Qwen3-32B partition reproduces on the pinned revision.

Model or resource	Pinned revision
Qwen2.5-VL-7B-Instruct	`cc594898137f460bfe9f0759e9844b3ce807cfb5`
Qwen3-VL-32B-Instruct	`0cfaf48183f594c314753d30a4c4974bc75f3ccb`
Qwen2.5-VL-72B-Instruct	`89c86200743eec961a297729e7990e8f2ddbc4c5`
Qwen3-32B	`9216db5781bf21249d130ec9da846c4624c16137`
Llama-3.3-70B-Instruct	`6f6073b423013f6a7d4d9f39144961bfbfbc386b`
Qwen2.5-72B-Instruct	`495f39366efef23836d0cfae4fbe635880d2be31`
Emotion classifier	`0e1cd914e3d46199ed785853e12b57304e04178b`
Likelihood gate	`2290a62682d06624634c1f46a6ad5be0f47f38aa`
EmoBench-M media	`5659068f9573bad8dbe2fe86b4a40fc09030c2c7`
CREMA-D repository	`1658cd342dff90010aa843eaeebd53610a08b1dc`

The saved GPU environment records PyTorch 2.14.0, Transformers 4.57.6, and Accelerate 1.15.0. Inference uses bfloat16 and scaled dot-product attention. The external models are `j-hartmann/emotion-english-distilroberta-base` for emotion classification and `distilbert/distilgpt2` for the likelihood gate. Scoring occurs after generation without intervention hooks. The classifier's model card documents an English training mixture including MELD (Hartmann, 2022); we do not assume independence from every benchmark source.

A.2 Module discovery, DiffMean directions, and calibration

Discovery uses 901 emotional and topic-matched neutral texts from EmoBench-M; 301 further texts are held out for the ablation pilot. For each text, each feed-forward unit's input to the down projection is pooled as the mean absolute value over nonpadding tokens, and the resulting unit-by-text matrix is z-scored per unit. Our IterD-style partitioning, which differs from the original activation-modularity objective (Yu et al., 2026), alternates greedy reassignment of units and texts among $K = 5$ modules to maximize the bipartite modularity (Barber, 2007) of the rectified (ReLU) unit-by-text matrix times a size-balance term, minus $\lambda = 0.5$ times a hinge penalty ($\tau = 0.1$) on modules whose units do not prefer emotional over topic-matched neutral texts; it keeps the best of ten restarts of at most 25 sweeps with a minimum module size of 2%. The Qwen2.5-VL-7B and Qwen3-VL-32B partitions use these settings on their language towers with text-only inputs. The provenance check recomputes the same objective on freshly extracted activations for the stored Qwen3-32B, Llama-3.3-70B, and Qwen2.5-72B assignments; all three reproduce, with relative modularity differences of 1.50×10^{-7}, 3.43×10^{-8}, and 1.43×10^{-7} and module-contrast correlations of 1.00; their objective totals (-0.074, -0.050, and -0.090) exceed those of size-matched random relabelings

(-0.198, -0.167, and -0.198). The historical Qwen3-32B and Qwen3-VL-32B partitions did not record their discovery-time weight revision; the Qwen3-32B check establishes reproduction on the pinned weights, the Qwen3-VL-32B partition behind the group signatures was not re-verified, and disjointness between the historical vision-language discovery corpus and the new evaluation has not been established.

DiffMean directions use the same 901 discovery texts, pooled residual streams over nonpadding tokens, and the mean residual token norm of each layer for scaling. The development grid covers $c \in \{0.05, 0.1, 0.2, 0.4\}$ at layers one quarter, one half, and three quarters of depth, and module strengths $\alpha \in \{0.25, 0.5, 1, 2, 4, 8\}$; among settings whose mean development gate-pass rate is at least 80% of the baseline rate, the rule selects the one with the largest mean $\Delta\mathrm{SSR}$ across targets, with ties going to the smaller strength. The cross-modal test adds the direction at the development-selected layer with $c \in \{0.05, 0.1, 0.2, 0.4\}$, uses a softmax over the six label scores, tests actor-level sign flips (19,999 draws) with Holm correction over the 20 paired DiffMean$-$random tests and, for modules, over the 15 module conditions, and applies each VLM's own modules at $\alpha \in \{0.5, 1, 2\}$. The 32 development prompts are disjoint from the 128 evaluation prompts, the 32 calibration prompts, and the discovery texts. Development runs use one seed; the selected settings are module $\alpha = 1$ and DiffMean layer 48, $c = 0.4$ for Qwen3-32B; module $\alpha = 0.25$ and layer 40, $c = 0.1$ for Llama-3.3-70B; module $\alpha = 1$ and layer 60, $c = 0.4$ for Qwen2.5-72B; module $\alpha = 0.5$ and layer 21, $c = 0.2$ for Qwen2.5-VL-7B; and module $\alpha = 0.25$ and layer 48, $c = 0.4$ for Qwen3-VL-32B. In Qwen2.5-VL-7B the rule selected $c = 0.2$ at layer 21 because the only settings with larger development gains (layer 14 at $c = 0.2$ and 0.4, 3.65 points each) fell below the gate criterion. Content-free calibration averages the log-softmax label scores of the three content-free inputs for each condition; for T+I, the three transcripts are paired with the three flat images.

A.3 Families, units, and uncertainty

Table 2 lists the test unit and Holm family of each analysis.

Table 2: Statistical scope. Module steering families are $\alpha = 4$ (15), $\alpha \in \{2, 8\}$ (30), their union (45), $\alpha \leq 1$ (45), and all strengths (90). The CPU group-signature and quality analyses are post-hoc follow-ups; the added transfer, probe, label-prior, steering, and cross-modal analyses follow a protocol fixed before those experiments were run. All bootstrap intervals use 10,000 draws and are nominal 95% intervals.

Analysis	Test or interval unit	Holm family
Conversation T–I	Paired items; group-bootstrap intervals	3 models per rule
CREMA-D 3-frame gain	Paired clips; actor-bootstrap intervals	2 models
Centroid transfer	1,999 refitted reference-label shuffles	2 models (primary)
Probes	1,999 within-group label shuffles; full procedure refit	15 original / 20
Module steering	19,999 prompt sign flips; seeds averaged	15 / 30 / 45 / 45 / 90
Matched evaluation	Same paired prompt design	17 per model (5)
Cross-modal intervention test	19,999 actor sign flips	20 DiffMean$-$random; 15 modules (per VLM)
Ablation pilot	5 matched random unit sets per module	Raw p only

Add-one permutation p values use $(1+\#\{s_b \geq s_{\mathrm{obs}}\})/(B+1)$ for one-sided transfer and probe tests. The steering statistic is the mean paired target-and-gate difference after averaging the three seeds within each prompt; its two-sided sign-flip test compares absolute statistics. Exact class-permutation representational similarity analysis (RSA; Kriegeskorte et al., 2008) includes all 720 permutations and uses the two-sided absolute correlation. Bootstrap resampling retains records within the selected group, so actor and dialogue dependence are partly represented in intervals. Tests and intervals are not interchangeable, and intervals excluding zero need not match corrected test decisions. The power simulation holds each model's baseline records fixed, draws steered indicators as independent Bernoulli variables with the baseline rate plus a gain, and applies the same two-sided prompt sign-flip statistic with 4,999 flips (400 simulations per gain).

B Complete Behavioral Results

Table 3: Conversational accuracy (%, $n = 1{,}411$) under raw label scoring. Δ is text minus frame accuracy in percentage points; intervals resample groups. Holm p covers three paired T–I tests. T+I receives both inputs. Calibrated results are in Table 5.

Model	Text	Frame	Both	Δ [95% CI]	Holm p
Qwen2.5-VL-7B	31.47	26.65	33.31	4.82 [1.68, 7.97]	0.0017
Qwen3-VL-32B	39.33	26.65	37.92	12.69 [9.31, 16.03]	9.56×10^{-14}
Qwen2.5-VL-72B	38.34	25.73	37.92	12.62 [9.41, 15.74]	7.20×10^{-15}

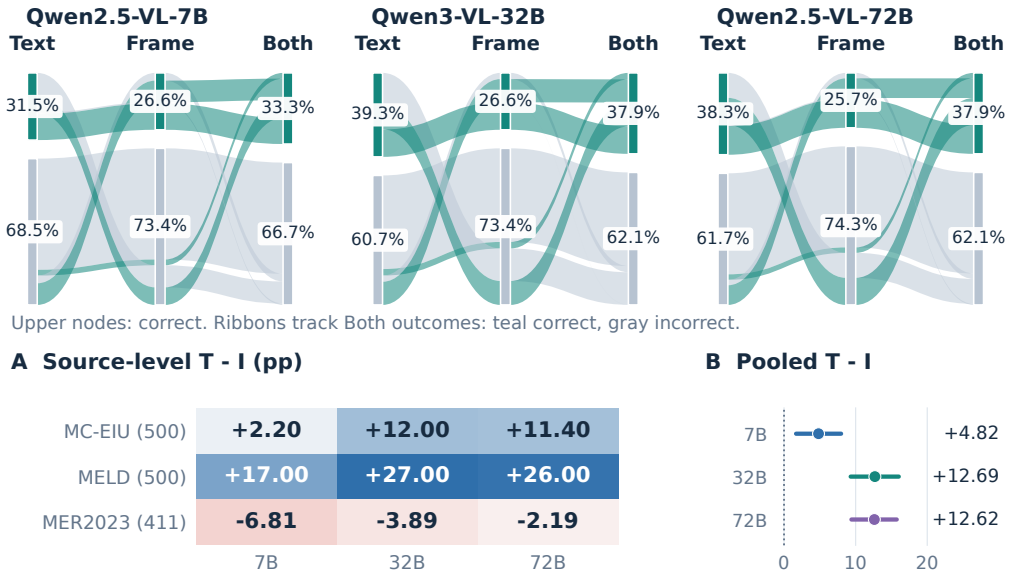

Figure 4: Paired correctness across input conditions ($n = 1{,}411$ per model). Ribbons preserve complete Text/Frame/Both correctness paths; their widths give item counts, upper nodes are correct, and color tracks the Both outcome. These are matched input conditions, not temporal transitions. (A) Descriptive text-minus-frame differences by source, in percentage points. (B) Pooled differences with marginal 95% dialogue/video-group bootstrap intervals. All three paired T–I tests pass Holm correction.

The conversational label counts are 252 anger, 110 disgust, 154 fear, 238 joy, 255 neutral, 213 sadness, and 189 surprise. The equal frame counts of Qwen2.5-VL-7B and Qwen3-VL-32B (376 each) are coincidental: Qwen2.5-VL-7B predicts neutral for 907 frames, whereas Qwen3-VL-32B predicts surprise for 510 (Table 6). Content-free calibration eliminates neutral frame predictions for Qwen3-VL-32B and Qwen2.5-VL-72B. The MER2023 frame advantage is 9.00, 7.79, and 6.57 points after label-free prior removal; after content-free calibration it is 14.84 and 1.46 points for Qwen2.5-VL-7B and Qwen2.5-VL-72B, whereas Qwen3-VL-32B shows a 0.73-point transcript advantage. Item-level annotation could separate the candidate explanations for this source reversal. No balancing weights are applied to the pooled accuracy; Table 7 gives macro-F1, and Table 8 repeats the comparison on a 66-item subset. Source groups preserve MELD dialogues, while MC-EIU and MER2023 use video identifiers; shared identities or scenes across sources can remain unaccounted for.

For CREMA-D, 7B midpoint and three-frame recognition yield 79/240 and 84/240 correct predictions under raw scoring; the gain is 2.08 points, with actor interval $[-0.84, 4.94]$ and adjusted $p = 0.267$. The 32B values are 137/240 and 149/240, a 5.00-point gain with actor interval [0.84, 9.42] and adjusted $p = 0.100$; the marginal interval excludes zero although the corrected test is not significant,

Table 4: Descriptive source-level accuracy (%) and T–I difference (percentage points) under raw scoring. These are not independent source-level significance tests. The same items are reused across models and conditions.

Model	Source	n	Text	Frame	Both	Δ (pp)
Qwen2.5-VL-7B	MC-EIU	500	26.80	24.60	29.20	+2.20
Qwen2.5-VL-7B	MELD	500	36.80	19.80	31.00	+17.00
Qwen2.5-VL-7B	MER2023	411	30.66	37.47	41.12	−6.81
Qwen3-VL-32B	MC-EIU	500	32.80	20.80	32.40	+12.00
Qwen3-VL-32B	MELD	500	49.20	22.20	38.80	+27.00
Qwen3-VL-32B	MER2023	411	35.28	39.17	43.55	−3.89
Qwen2.5-VL-72B	MC-EIU	500	33.00	21.60	31.20	+11.40
Qwen2.5-VL-72B	MELD	500	46.40	20.40	36.80	+26.00
Qwen2.5-VL-72B	MER2023	411	35.04	37.23	47.45	−2.19

Table 5: Conversational accuracy (%) under raw scoring, label-free prior removal, and content-free calibration. T−I is in percentage points with a group-bootstrap interval; Holm p covers three models within each rule. The last column is the descriptive MER2023 T−I difference in percentage points; negative values favor frames.

Model	Rule	Text	Frame	Both	T−I [95% CI]	Holm p	MER2023 T−I
Qwen2.5-VL-7B	Raw	31.47	26.65	33.31	4.82 [1.68, 7.97]	0.0017	−6.81
Qwen2.5-VL-7B	Prior removed	33.95	26.51	33.66	7.44 [4.02, 10.72]	1.47×10^{-5}	−9.00
Qwen2.5-VL-7B	Content-free	33.59	27.29	34.59	6.31 [2.75, 9.85]	3.39×10^{-4}	−14.84
Qwen3-VL-32B	Raw	39.33	26.65	37.92	12.69 [9.31, 16.03]	9.56×10^{-14}	−3.89
Qwen3-VL-32B	Prior removed	39.19	29.55	41.60	9.64 [6.23, 13.07]	3.02×10^{-8}	−7.79
Qwen3-VL-32B	Content-free	37.56	23.60	32.67	13.96 [10.92, 17.01]	2.13×10^{-18}	0.73
Qwen2.5-VL-72B	Raw	38.34	25.73	37.92	12.62 [9.41, 15.74]	7.20×10^{-15}	−2.19
Qwen2.5-VL-72B	Prior removed	38.77	27.92	39.62	10.84 [7.33, 14.19]	3.75×10^{-10}	−6.57
Qwen2.5-VL-72B	Content-free	36.29	24.03	34.80	12.26 [9.07, 15.43]	7.64×10^{-15}	−1.46

Table 6: Predicted-label counts on the 1,411 conversational items for each model, input condition, and scoring rule. Gold counts are 252 anger, 110 disgust, 154 fear, 238 joy, 255 neutral, 213 sadness, and 189 surprise. Label-free prior removal spreads frame predictions across labels in all three models; content-free calibration removes the neutral mode of Qwen2.5-VL-7B and Qwen2.5-VL-72B but concentrates Qwen3-VL-32B frame predictions further on surprise (787 of 1,411).

Model	Rule	Input	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Qwen2.5-VL-7B	Raw	T	100	26	65	354	556	237	73
Qwen2.5-VL-7B	Raw	I	32	1	9	199	907	187	76
Qwen2.5-VL-7B	Raw	T+I	59	9	27	293	626	318	79
Qwen2.5-VL-7B	Prior removed	T	175	158	228	213	306	123	208
Qwen2.5-VL-7B	Prior removed	I	213	104	172	226	191	247	258
Qwen2.5-VL-7B	Prior removed	T+I	199	144	203	238	172	221	234
Qwen2.5-VL-7B	Content-free	T	127	288	153	186	215	198	244
Qwen2.5-VL-7B	Content-free	I	166	50	29	217	331	269	349
Qwen2.5-VL-7B	Content-free	T+I	217	112	77	232	192	322	259
Qwen3-VL-32B	Raw	T	257	67	88	269	178	237	315
Qwen3-VL-32B	Raw	I	106	62	32	222	101	378	510
Qwen3-VL-32B	Raw	T+I	159	70	52	209	86	340	495
Qwen3-VL-32B	Prior removed	T	222	101	140	279	295	155	219
Qwen3-VL-32B	Prior removed	I	163	88	104	305	235	225	291
Qwen3-VL-32B	Prior removed	T+I	189	98	146	300	233	203	242
Qwen3-VL-32B	Content-free	T	241	129	138	299	54	123	427
Qwen3-VL-32B	Content-free	I	64	65	58	217	0	220	787
Qwen3-VL-32B	Content-free	T+I	127	59	76	185	0	198	766
Qwen2.5-VL-72B	Raw	T	184	35	91	182	582	124	213
Qwen2.5-VL-72B	Raw	I	51	0	13	149	824	69	305
Qwen2.5-VL-72B	Raw	T+I	137	17	47	185	558	132	335
Qwen2.5-VL-72B	Prior removed	T	224	105	210	320	209	174	169
Qwen2.5-VL-72B	Prior removed	I	219	96	159	351	102	222	262
Qwen2.5-VL-72B	Prior removed	T+I	200	110	219	379	94	215	194
Qwen2.5-VL-72B	Content-free	T	92	226	140	317	7	311	318
Qwen2.5-VL-72B	Content-free	I	38	313	95	254	0	358	353
Qwen2.5-VL-72B	Content-free	T+I	70	261	134	283	0	392	271

Model	Text	Frame	Both
Qwen2.5-VL-7B	29.16	18.28	27.42
Qwen3-VL-32B	37.90	22.87	35.68
Qwen2.5-VL-72B	37.38	19.07	35.59

Table 7: Seven-label macro-F1 (%) on the 1,411-item corpus under raw scoring. Accuracy and macro-F1 differ, particularly for image-only inputs, so accuracy should not be read as balanced performance across emotions.

Table 8: Fresh midpoint-frame evaluation of the overlapping 66-item subset, in %. It contains 29 MC-EIU, 17 MELD, and 20 MER2023 items and is not an independent replication. None of its three paired T–I tests survives Holm correction. Historical thumbnails and predictions are not pooled with these runs.

Model	Text	Frame	Both	Δ [95% CI]	Holm p
Qwen2.5-VL-7B	37.88	28.79	30.30	9.09 [−4.55, 22.73]	0.263
Qwen3-VL-32B	42.42	24.24	40.91	18.18 [3.03, 33.33]	0.072
Qwen2.5-VL-72B	46.97	27.27	43.94	19.70 [4.55, 34.85]	0.072

Table 9: CREMA-D comparison under raw scoring and the text-mean transfer readout (%). Reference-text classification uses 120 balanced exemplars; frame recognition and transfer use 240 clips. Transfer intervals resample 88 actors, and Holm p covers that analysis's four-test family of two transfer and two geometry tests. The corrected readouts are in Table 10.

Model	Ref. text	Frame	3 frames	Transfer [95% CI]	Holm p
Qwen2.5-VL-7B	35.83	32.92	35.00	27.50 [23.48, 31.56]	0.0060
Qwen3-VL-32B	41.67	57.08	62.08	16.67 [13.28, 20.09]	0.930

Table 10: CREMA-D frame recognition under the three scoring rules and text-centroid transfer under the three readouts (%). Transfer p values use 1,999 refitted reference-label permutations; Holm p covers the prespecified modality-centered readout in two models. The last column gives the number of classes each readout predicts.

	Frame accuracy			Text-centroid transfer to single frames						
Model	Raw	Prior rem.	Content-free	Text-mean	p	Centered [95% CI]	Holm p	z-scored	p	Classes
Qwen2.5-VL-7B	32.92	51.67	42.08	27.50	0.0015	32.92 [27.54, 38.60]	0.0040	37.50	0.0015	3/6/6
Qwen3-VL-32B	57.08	57.08	53.33	16.67	0.930	21.25 [16.60, 26.16]	0.116	43.75	5.00×10^{-4}	1/6/6

and we retain the corrected decision. Reference-text classification yields 43/120 and 50/120 correct predictions (Table 9). These texts are category exemplars, not lexical matches to CREMA-D clips.

C Probe and Class-Geometry Details

Table 11: Held-out conversational text-to-image probes ($n = 1{,}411$). Accuracy and conditional null means (1,999 within-group permutations) are percentages. "Full" denotes mean final-layer features; 6 and 384 denote group-signature dimensions. The text scaler is fitted on training-fold text; the image scaler uses unlabeled image training-fold statistics. Holm-20 corrects all 20 probe tests (Table 12); the last column tests accuracy against the 18.07% majority rate.

Model	Features	Transfer	Acc.	Null	Raw p	Holm-20 p	p vs majority
Qwen2.5-VL-7B	Full	Text scaler	17.86	15.35	0.045	0.540	0.593
Qwen2.5-VL-7B	Full	Image scaler	20.06	18.79	0.129	1.000	0.030
Qwen3-VL-32B	Full	Text scaler	16.51	12.20	0.0045	0.068	0.942
Qwen3-VL-32B	Full	Image scaler	19.49	18.51	0.232	1.000	0.090
Qwen2.5-VL-72B	Full	Text scaler	16.51	16.32	0.494	1.000	0.942
Qwen2.5-VL-72B	Full	Image scaler	17.43	17.80	0.637	1.000	0.743
Qwen3-VL-32B	6	Text scaler	17.08	16.89	0.235	1.000	0.842
Qwen3-VL-32B	6	Image scaler	18.07	17.77	0.295	1.000	0.511
Qwen3-VL-32B	384	Text scaler	12.97	10.73	0.209	1.000	1.000
Qwen3-VL-32B	384	Image scaler	15.24	16.12	0.810	1.000	0.998

Figure 5 compares every probe score with the mean of its conditional null. Null means differ across representations and tasks because the shuffles preserve group label composition, so a score above its null mean is not necessarily above the majority rate. The strongest text-scaler transfer (16.51% against a 12.20% null mean; raw $p = 0.0045$) has Holm $p = 0.045$ in the original 15-test family and 0.068 over all 20 tests. Standardizing images with their own statistics raises full-feature transfer to 17.43–20.06%, but none of these exceeds its null after correction.

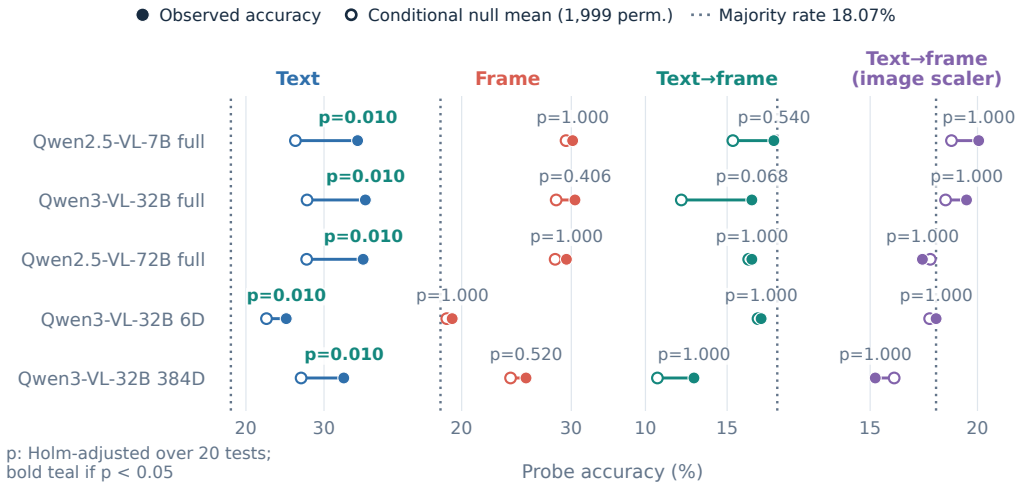

Figure 5: All 20 probe conditions on 1,411 conversational items. Filled points show observed accuracy and open points the mean of 1,999 refitted within-group permutations; the dotted line is the 18.07% majority rate. Annotations give Holm p over all 20 tests. Only 401 labels can move.

Group signatures have shape $1{,}411 \times 64 \times 6$. Layer averaging produces six features and layer-major flattening produces 384 features, without fitting an image–text map. Each shuffled label vector receives its own stratified group fold construction; all training folds contain all seven classes and no group crosses train/test. The 6D transfer accuracy is 241/1,411 (17.08%; fixed-prediction group interval [15.04, 19.14]%); the 384D transfer is 183/1,411 (12.97%; [11.07, 14.87]%). Their raw conditional p values are 0.235 and 0.209. Passing this null only establishes performance beyond within-group label reassignment; it does not establish speaker-independent or source-independent

Table 12: Every conversational probe condition with 1,999 conditional permutations. Acc. and Null are observed accuracy and conditional null mean (%). Holm-15 adjusts the 15 original tests; Holm-20 adds the five image-scaler transfer tests. T and I denote within-modality decoding; T→I trains on text and tests held-out images.

Model	Features	Task	Acc.	Null	Raw p	Holm-15 p	Holm-20 p
Qwen2.5-VL-7B	Full	T	34.30	26.34	5.00×10^{-4}	0.0075	0.010
Qwen2.5-VL-7B	Full	I	30.12	29.52	0.260	1.000	1.000
Qwen2.5-VL-7B	Full	T→I	17.86	15.35	0.045	0.320	0.540
Qwen2.5-VL-7B	Full	T→I (img. scaler)	20.06	18.79	0.129	–	1.000
Qwen3-VL-32B	Full	T	35.29	27.83	5.00×10^{-4}	0.0075	0.010
Qwen3-VL-32B	Full	I	30.33	28.62	0.029	0.261	0.406
Qwen3-VL-32B	Full	T→I	16.51	12.20	0.0045	0.045	0.068
Qwen3-VL-32B	Full	T→I (img. scaler)	19.49	18.51	0.232	–	1.000
Qwen2.5-VL-72B	Full	T	35.01	27.78	5.00×10^{-4}	0.0075	0.010
Qwen2.5-VL-72B	Full	I	29.55	28.53	0.114	0.681	1.000
Qwen2.5-VL-72B	Full	T→I	16.51	16.32	0.494	1.000	1.000
Qwen2.5-VL-72B	Full	T→I (img. scaler)	17.43	17.80	0.637	–	1.000
Qwen3-VL-32B	6	T	25.16	22.63	5.00×10^{-4}	0.0075	0.010
Qwen3-VL-32B	6	I	19.14	18.64	0.217	1.000	1.000
Qwen3-VL-32B	6	T→I	17.08	16.89	0.235	1.000	1.000
Qwen3-VL-32B	6	T→I (img. scaler)	18.07	17.77	0.295	–	1.000
Qwen3-VL-32B	384	T	32.53	27.07	5.00×10^{-4}	0.0075	0.010
Qwen3-VL-32B	384	I	25.87	24.46	0.040	0.320	0.520
Qwen3-VL-32B	384	T→I	12.97	10.73	0.209	1.000	1.000
Qwen3-VL-32B	384	T→I (img. scaler)	15.24	16.12	0.810	–	1.000

generalization. The first 199 of the 1,999 null draws are those reported in an earlier analysis, under which no pooled test could reach 0.05.

For CREMA-D under the text-mean readout, the 7B transfer null averages 16.67% and the 32B null 16.69%; 85.29% of the 32B permutations also yield exactly 16.67% accuracy, so that null is nearly degenerate. The geometry comparison uses image labels to form class centroids, unlike the transfer predictor, whose fitting uses reference-text labels only. It is a descriptive class-structure test; representational similarity need not track functional behavior (Ding et al., 2021), so it does not identify an emotion-specific functional subspace.

C.1 Descriptive layer-by-group fingerprints

To visualize the saved signatures without fitting another readout, we compute a class-versus-rest contrast for modality v, emotion c, layer ℓ, and group g:

$$D^{(v)}_{c\ell g} = \frac{\overline{s}^{(v)}_{c\ell g} - \overline{s}^{(v)}_{\neg c,\ell g}}{\mathrm{SD}_i(s^{(v)}_{i\ell g})}. \tag{4}$$

The denominator is the population standard deviation over all 1,411 items within the same modality. At each layer, we compare the two vectors of $7 \times 6 = 42$ emotion/group contrasts by cosine similarity. Figure 6 shows the text, frame, and difference maps on a common scale. These post-hoc summaries use all labels and the fixed historical partition, with no held-out fitting or inferential test, and visible structure does not establish a transferable decoder or a shared causal circuit.

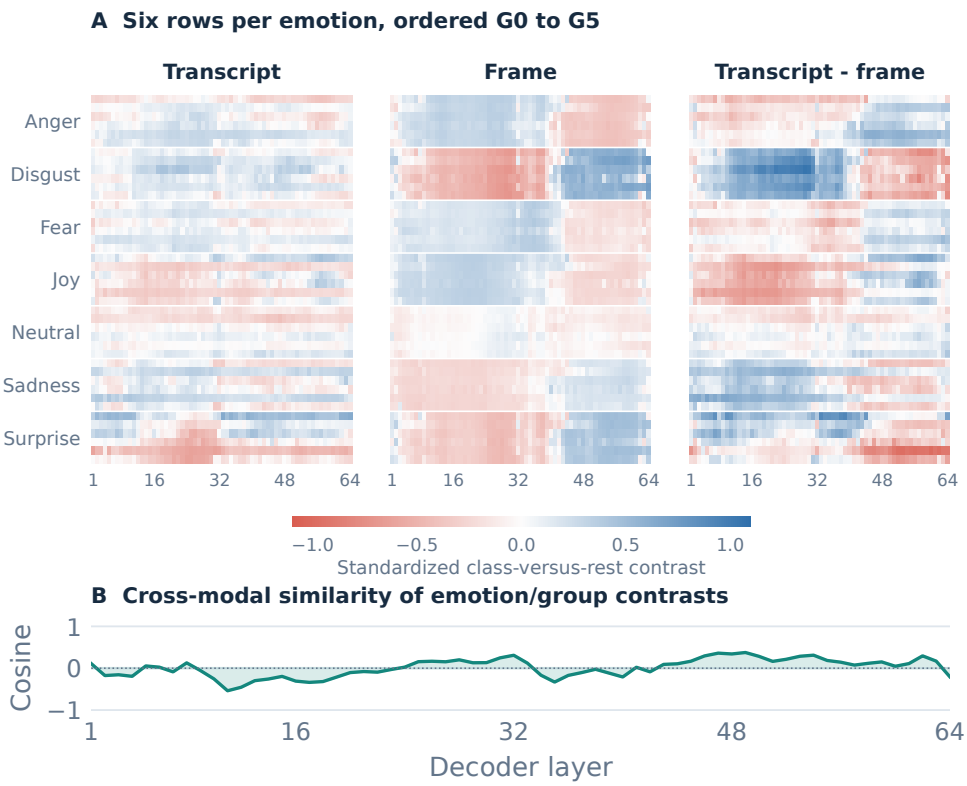

Figure 6: Descriptive Qwen3-VL-32B activation fingerprints from 1,411 paired items. (A) Standardized class-versus-rest contrasts across 64 layers, with six rows per emotion ordered G0–G5. Blue is positive and coral negative; transcript, frame, and their difference share a symmetric scale covering all observed values. (B) Layerwise cosine similarity between the 42-dimensional text and frame contrast vectors. These all-label summaries are distinct from the held-out probes and do not localize an emotion circuit.

D All Steering Conditions

The baseline likelihood-gate pass rates in the historical runs are 97.14% for Qwen3-32B, 95.57% for Llama-3.3-70B, and 95.83% for Qwen2.5-72B. The historical primary and sensitivity runs reuse the same deterministic baseline and calibration outputs; their saved outputs comprise 18,432 unique evaluation outputs (1,152 baseline and 17,280 intervention) and 96 calibration outputs. These are repeated generations of 128 evaluation prompts per model, not thousands of independent prompts. The prompt requests one or two continuation sentences without an explicit emotion instruction. The 128 evaluation, 32 calibration, and 32 development prompts wrap distinct sentiment-neutral transcripts from the MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), CH-SIMS v2 (Liu et al., 2022), and FMSA-SC (Song et al., 2024) subsets of EmoBench-M (Hu et al., 2025), which do not overlap the three recognition sources. Unit scales are estimated from the first 64 nonempty discovery texts; the separate 32-prompt calibration set determines only the gate threshold. Seeds are paired between baseline and intervention. Nine of the 15 historical module targets are anger, three sadness, two joy, and one surprise, whereas DiffMean covers all six non-neutral targets. In the matched evaluation, DiffMean steering lowers the gate-pass rate for four of six Qwen3-32B targets (81.51–92.97%, against a 97.14% baseline). In the Qwen2.5-VL-7B cross-modal test, the anger direction has a smaller effect that exceeds the random direction (1.45 points; 3.22 above random, Holm $p = 0.001$) but not the off-target shifts (margin -4.25 [-5.58, -2.97] points), the disgust direction does not exceed the random direction, the fear direction does not differ from it, and the sadness direction lowers the probability of sadness (Table 20). Gate is the saved proxy pass rate and Spec. is Equation 3; neither is a human quality score. Figure 7 and Tables 13 and 14 report the 45 historical conditions, and Table 15 the 45 low-strength conditions.

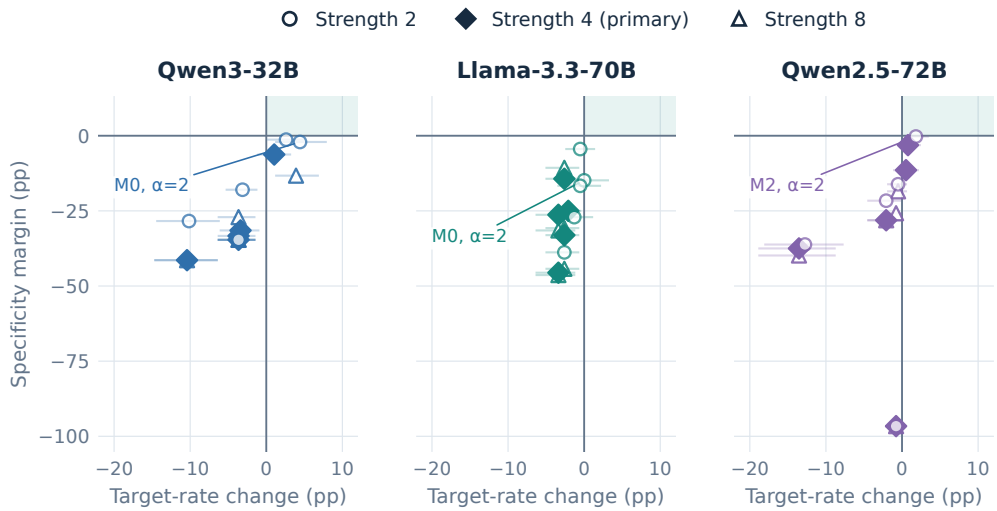

Figure 7: The 45 historical module conditions in target-gain/specificity space. Each point uses 128 prompts and three paired seeds; shapes distinguish strengths and filled diamonds mark $\alpha = 4$. Horizontal intervals are nominal 95% prompt-bootstrap intervals for ΔSSR; specificity is the ungated margin in Equation 3. The shaded quadrant combines positive target gain and specificity. All margins are negative, and no positive gain passes correction. Annotations identify the largest ΔSSR estimate per model, which is zero for Llama-3.3-70B.

Table 13: All 15 historical primary conditions at $\alpha = 4$. ΔSSR, its nominal prompt-bootstrap interval, gate-pass rate, and specificity margin are in percentage points or percent. Holm p adjusts all 15 primary tests across models.

Model	Mod.	α	Target	ΔSSR [95% CI]	Gate	Spec.	Holm p
Qwen3-32B	M0	4	anger	−3.65 [−6.25, −1.56]	90.63	−33.33	0.025
Qwen3-32B	M1	4	surprise	1.04 [−1.04, 3.13]	94.53	−6.25	1.000
Qwen3-32B	M2	4	joy	−10.42 [−14.58, −6.51]	97.14	−41.41	7.50×10^{-4}
Qwen3-32B	M3	4	anger	−3.65 [−6.25, −1.56]	0.00	−34.64	0.025
Qwen3-32B	M4	4	anger	−3.39 [−5.99, −1.04]	86.72	−31.51	0.094
Llama-3.3-70B	M0	4	sadness	−3.39 [−6.25, −1.30]	13.02	−26.30	0.094
Llama-3.3-70B	M1	4	anger	−2.60 [−4.95, −0.78]	0.00	−33.07	0.124
Llama-3.3-70B	M2	4	sadness	−3.39 [−6.25, −1.30]	0.00	−45.57	0.094
Llama-3.3-70B	M3	4	anger	−2.08 [−3.91, −0.52]	92.19	−25.00	0.190
Llama-3.3-70B	M4	4	anger	−2.60 [−4.95, −0.78]	0.00	−14.32	0.124
Qwen2.5-72B	M0	4	anger	0.52 [−1.04, 2.08]	66.93	−11.46	1.000
Qwen2.5-72B	M1	4	sadness	−2.08 [−4.43, −0.26]	0.00	−28.13	0.311
Qwen2.5-72B	M2	4	anger	0.78 [−0.26, 2.34]	96.35	−3.13	1.000
Qwen2.5-72B	M3	4	anger	−0.78 [−2.08, 0.00]	100.00	−96.61	1.000
Qwen2.5-72B	M4	4	joy	−13.54 [−18.75, −8.85]	1.82	−37.50	7.50×10^{-4}

Table 14: All 30 historical sensitivity conditions at $\alpha \in \{2, 8\}$. Units match Table 13; Holm p adjusts the 30 sensitivity tests across models.

Model	Mod.	α	Target	ΔSSR [95% CI]	Gate	Spec.	Holm p
Qwen3-32B	M0	2	anger	4.43 [1.30, 7.81]	79.43	−2.08	0.280
Qwen3-32B	M0	8	anger	−3.65 [−6.25, −1.56]	29.69	−34.64	0.051
Qwen3-32B	M1	2	surprise	2.60 [0.26, 4.95]	85.94	−1.30	0.606
Qwen3-32B	M1	8	surprise	3.91 [1.30, 6.77]	99.74	−13.28	0.180
Qwen3-32B	M2	2	joy	−10.16 [−14.33, −6.25]	89.06	−28.39	0.0015
Qwen3-32B	M2	8	joy	−10.42 [−14.58, −6.51]	100.00	−41.41	0.0015
Qwen3-32B	M3	2	anger	−3.65 [−6.25, −1.56]	100.00	−34.64	0.051
Qwen3-32B	M3	8	anger	−3.65 [−6.25, −1.56]	7.55	−34.64	0.051
Qwen3-32B	M4	2	anger	−3.13 [−5.21, −1.30]	61.98	−17.97	0.051
Qwen3-32B	M4	8	anger	−3.65 [−6.25, −1.56]	4.43	−27.08	0.051
Llama-3.3-70B	M0	2	sadness	0.00 [−3.39, 3.13]	100.00	−14.84	1.000
Llama-3.3-70B	M0	8	sadness	−3.39 [−6.25, −1.30]	0.00	−31.51	0.179
Llama-3.3-70B	M1	2	anger	−2.60 [−4.95, −0.78]	0.00	−38.80	0.280
Llama-3.3-70B	M1	8	anger	−2.60 [−4.95, −0.78]	0.00	−10.68	0.280
Llama-3.3-70B	M2	2	sadness	−0.52 [−3.39, 2.08]	65.89	−16.67	1.000
Llama-3.3-70B	M2	8	sadness	−3.39 [−6.25, −1.30]	0.00	−46.35	0.179
Llama-3.3-70B	M3	2	anger	−0.52 [−2.34, 1.30]	98.44	−4.43	1.000
Llama-3.3-70B	M3	8	anger	−2.60 [−4.95, −0.78]	0.00	−30.73	0.280
Llama-3.3-70B	M4	2	anger	−1.30 [−3.65, 1.04]	95.05	−27.08	1.000
Llama-3.3-70B	M4	8	anger	−2.60 [−4.95, −0.78]	0.00	−44.27	0.280
Qwen2.5-72B	M0	2	anger	−0.52 [−1.82, 0.52]	99.74	−16.15	1.000
Qwen2.5-72B	M0	8	anger	−0.78 [−2.08, 0.00]	2.86	−25.78	1.000
Qwen2.5-72B	M1	2	sadness	−2.08 [−4.43, −0.26]	97.92	−21.61	0.684
Qwen2.5-72B	M1	8	sadness	−2.08 [−4.43, −0.26]	18.49	−28.13	0.684
Qwen2.5-72B	M2	2	anger	1.82 [0.52, 3.39]	97.92	−0.26	0.407
Qwen2.5-72B	M2	8	anger	−0.52 [−1.82, 0.52]	98.44	−18.49	1.000
Qwen2.5-72B	M3	2	anger	−0.78 [−2.08, 0.00]	100.00	−96.61	1.000
Qwen2.5-72B	M3	8	anger	−0.78 [−2.08, 0.00]	100.00	−96.61	1.000
Qwen2.5-72B	M4	2	joy	−12.76 [−17.97, −7.81]	97.14	−36.20	0.0015
Qwen2.5-72B	M4	8	joy	−13.54 [−18.75, −8.85]	5.47	−39.84	0.0015

Table 15: All 45 low-strength module conditions ($\alpha \leq 1$) for the three text models, with prompt-bootstrap intervals for ΔSSR and the specificity margin, both in percentage points; Gate is the steered gate-pass rate (%). Holm-45 adjusts these 45 tests across models. The rows at each model's development-selected strength ($\alpha = 1$ for the Qwen models, $\alpha = 0.25$ for Llama-3.3-70B) repeat the module rows of Table 18, where a different family applies.

Model	Mod.	α	Target	ΔSSR [95% CI]	Gate	Spec. [95% CI]	Holm-45 p
Qwen3-32B	M0	0.25	anger	−2.34 [−4.43, −0.52]	94.79	−4.17 [−9.64, −2.08]	1.000
Qwen3-32B	M0	0.5	anger	−0.78 [−2.86, 1.30]	94.53	−1.82 [−7.03, 0.00]	1.000
Qwen3-32B	M0	1	anger	0.00 [−2.35, 2.60]	93.75	−2.34 [−9.11, 1.30]	1.000
Qwen3-32B	M1	0.25	surprise	1.30 [−0.52, 3.39]	95.05	0.00 [−3.39, 2.08]	1.000
Qwen3-32B	M1	0.5	surprise	1.30 [−0.52, 3.39]	92.71	0.00 [−3.91, 1.56]	1.000
Qwen3-32B	M1	1	surprise	2.08 [−0.52, 4.95]	90.36	0.00 [−4.17, 2.86]	1.000
Qwen3-32B	M2	0.25	joy	0.52 [−2.34, 3.65]	95.31	−0.26 [−5.73, 2.86]	1.000
Qwen3-32B	M2	0.5	joy	−1.56 [−4.95, 1.82]	96.88	−3.39 [−9.38, 0.26]	1.000
Qwen3-32B	M2	1	joy	−6.51 [−10.68, −2.60]	95.83	−14.58 [−23.44, −8.33]	0.101
Qwen3-32B	M3	0.25	anger	−3.65 [−6.25, −1.56]	99.74	−33.33 [−40.89, −26.30]	0.088
Qwen3-32B	M3	0.5	anger	−3.65 [−6.25, −1.56]	100.00	−101.56 [−104.43, −98.96]	0.088
Qwen3-32B	M3	1	anger	−3.65 [−6.25, −1.56]	94.27	−34.64 [−42.19, −27.34]	0.088
Qwen3-32B	M4	0.25	anger	−1.56 [−3.39, 0.00]	94.79	−5.21 [−11.20, −2.34]	1.000
Qwen3-32B	M4	0.5	anger	−1.30 [−3.39, 0.78]	95.57	−5.99 [−13.02, −2.60]	1.000
Qwen3-32B	M4	1	anger	−2.60 [−4.43, −0.78]	83.07	−14.06 [−19.01, −9.64]	0.449
Llama-3.3-70B	M0	0.25	sadness	−0.78 [−3.39, 1.56]	95.31	−3.91 [−9.38, −0.52]	1.000
Llama-3.3-70B	M0	0.5	sadness	−0.52 [−2.86, 1.56]	95.57	−4.17 [−9.64, −0.26]	1.000
Llama-3.3-70B	M0	1	sadness	0.78 [−1.82, 3.13]	98.44	−0.78 [−7.03, 1.82]	1.000
Llama-3.3-70B	M1	0.25	anger	−2.60 [−4.95, −0.78]	0.00	−45.83 [−53.91, −38.02]	0.622
Llama-3.3-70B	M1	0.5	anger	−2.60 [−4.95, −0.78]	0.00	−87.76 [−92.71, −82.55]	0.622
Llama-3.3-70B	M1	1	anger	−2.60 [−4.95, −0.78]	0.00	−39.84 [−46.88, −32.55]	0.622
Llama-3.3-70B	M2	0.25	sadness	−0.78 [−2.35, 0.52]	96.61	−1.56 [−5.73, −0.52]	1.000
Llama-3.3-70B	M2	0.5	sadness	−0.52 [−2.08, 0.78]	97.14	−2.08 [−7.29, −0.78]	1.000
Llama-3.3-70B	M2	1	sadness	−0.52 [−2.08, 1.04]	98.18	−2.08 [−7.55, −0.52]	1.000
Llama-3.3-70B	M3	0.25	anger	0.00 [−1.04, 1.04]	96.09	−0.78 [−4.69, 0.00]	1.000
Llama-3.3-70B	M3	0.5	anger	2.08 [0.26, 4.43]	96.09	1.04 [−2.86, 3.13]	1.000
Llama-3.3-70B	M3	1	anger	0.52 [−1.30, 2.08]	97.14	−0.52 [−4.69, 1.30]	1.000
Llama-3.3-70B	M4	0.25	anger	0.00 [−1.56, 1.56]	97.40	−3.39 [−8.33, −0.52]	1.000
Llama-3.3-70B	M4	0.5	anger	0.26 [−2.08, 2.60]	98.18	−5.21 [−11.72, −0.26]	1.000
Llama-3.3-70B	M4	1	anger	−1.30 [−3.65, 1.04]	95.05	−18.75 [−26.82, −10.68]	1.000
Qwen2.5-72B	M0	0.25	anger	−0.26 [−0.78, 0.00]	97.92	−2.34 [−5.99, −1.04]	1.000
Qwen2.5-72B	M0	0.5	anger	1.04 [0.26, 2.08]	95.31	−2.08 [−7.03, 0.52]	1.000
Qwen2.5-72B	M0	1	anger	0.26 [−0.52, 1.30]	94.79	−2.60 [−9.11, −0.52]	1.000
Qwen2.5-72B	M1	0.25	sadness	−0.52 [−2.08, 1.04]	96.35	−1.82 [−5.73, 0.00]	1.000
Qwen2.5-72B	M1	0.5	sadness	−0.52 [−2.08, 0.78]	96.09	−1.04 [−5.47, 0.01]	1.000
Qwen2.5-72B	M1	1	sadness	−0.78 [−2.60, 1.04]	94.27	−2.60 [−7.81, −0.78]	1.000
Qwen2.5-72B	M2	0.25	anger	0.00 [0.00, 0.00]	95.83	−0.78 [−2.60, −0.52]	1.000
Qwen2.5-72B	M2	0.5	anger	0.26 [0.00, 0.78]	96.35	−2.08 [−3.91, −0.78]	1.000
Qwen2.5-72B	M2	1	anger	0.78 [0.00, 1.82]	97.14	−1.56 [−6.51, 0.52]	1.000
Qwen2.5-72B	M3	0.25	anger	−0.78 [−2.08, 0.00]	100.00	−96.61 [−99.48, −93.49]	1.000
Qwen2.5-72B	M3	0.5	anger	−0.78 [−2.08, 0.00]	100.00	−96.61 [−99.48, −93.49]	1.000
Qwen2.5-72B	M3	1	anger	−0.78 [−2.08, 0.00]	100.00	−96.61 [−99.48, −93.49]	1.000
Qwen2.5-72B	M4	0.25	joy	1.56 [−1.56, 4.95]	97.14	0.26 [−4.43, 3.91]	1.000
Qwen2.5-72B	M4	0.5	joy	1.56 [−2.34, 5.73]	98.18	−1.30 [−7.29, 3.65]	1.000
Qwen2.5-72B	M4	1	joy	3.65 [−0.78, 8.07]	96.35	1.04 [−5.21, 5.73]	1.000

Table 16: Sources of the module decreases by strength. A gate-collapse floor has a steered gate-pass rate below 10%, which forces the gated target rate toward zero; the other decreases reflect a lower target rate among gate passes or gate loss without suppression. All five exactly zero estimates count as no decrease, including Llama-3.3-70B module 0 at $\alpha = 2$ and module 4 at $\alpha = 0.25$, which are stored as floating-point values just below zero. All three text models contribute at every strength.

α	No decrease	Gate floor	Suppressed	Gate loss
0.25	6	1	8	0
0.5	6	1	8	0
1	7	1	7	0
2	4	1	9	1
4	3	6	6	0
8	1	9	5	0

Table 17: Minimum detectable ΔSSR (MDE; percentage points) at 80% power for each historical target, at a nominal 0.05 level and at the Holm-worst level $0.05/15$, from 400 simulations per gain. The baseline gated target rate is in percent. Gains are searched on a grid (0.5, 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 12, 15, and 20 points), so each MDE is the smallest grid value reaching 80% power.

Model	Target	Baseline gated rate	MDE (level 0.05)	MDE (level $0.05/15$)
Qwen3-32B	anger	3.65	6.00	8.00
Qwen3-32B	joy	10.42	8.00	12.00
Qwen3-32B	surprise	1.82	4.00	6.00
Llama-3.3-70B	anger	2.60	5.00	8.00
Llama-3.3-70B	sadness	3.39	6.00	8.00
Qwen2.5-72B	anger	0.78	4.00	5.00
Qwen2.5-72B	joy	13.54	10.00	15.00
Qwen2.5-72B	sadness	2.08	5.00	8.00

Table 18: Matched evaluation in the three text models on 128 prompts × 3 seeds of the development-selected settings: DiffMean directions, norm-matched random directions, and modules. ΔSSR and Spec. are in percentage points with prompt-bootstrap intervals; Gate is in percent. Holm-17 adjusts the 17 conditions within each model. Table 19 gives the two vision-language models.

Model	Method	Setting	Target	ΔSSR [95% CI]	Gate	Spec. [95% CI]	Holm-17 p
Qwen3-32B	DiffMean	L=48, c=0.4	anger	10.68 [7.03, 14.84]	81.51	3.13 [−3.65, 7.81]	8.50×10^{-4}
Qwen3-32B	DiffMean	L=48, c=0.4	disgust	17.19 [13.02, 21.61]	98.18	5.99 [−0.78, 11.72]	8.50×10^{-4}
Qwen3-32B	DiffMean	L=48, c=0.4	fear	2.86 [−1.82, 7.55]	92.97	−8.07 [−15.36, −1.30]	1.000
Qwen3-32B	DiffMean	L=48, c=0.4	joy	55.73 [49.74, 61.46]	88.54	53.39 [46.09, 60.68]	8.50×10^{-4}
Qwen3-32B	DiffMean	L=48, c=0.4	sadness	2.34 [−0.26, 5.21]	98.70	0.26 [−3.65, 2.86]	0.981
Qwen3-32B	DiffMean	L=48, c=0.4	surprise	69.27 [64.32, 74.22]	90.63	76.56 [70.05, 81.26]	8.50×10^{-4}
Qwen3-32B	Random	L=48, c=0.4	anger	−3.39 [−5.99, −1.04]	80.99	−29.43 [−37.24, −21.88]	0.085
Qwen3-32B	Random	L=48, c=0.4	disgust	−0.52 [−1.82, 0.78]	92.71	−25.78 [−32.81, −19.01]	1.000
Qwen3-32B	Random	L=48, c=0.4	fear	−8.07 [−12.76, −3.91]	92.71	−22.40 [−32.29, −13.54]	0.0072
Qwen3-32B	Random	L=48, c=0.4	joy	−10.42 [−14.58, −6.51]	89.58	−28.91 [−35.16, −23.18]	8.50×10^{-4}
Qwen3-32B	Random	L=48, c=0.4	sadness	−1.04 [−2.86, 0.52]	54.17	−19.79 [−27.34, −12.24]	1.000
Qwen3-32B	Random	L=48, c=0.4	surprise	−0.26 [−2.08, 1.56]	93.75	−8.59 [−16.15, −3.91]	1.000
Qwen3-32B	Module M0	α=1	anger	0.00 [−2.35, 2.60]	93.75	−2.34 [−9.11, 1.30]	1.000
Qwen3-32B	Module M1	α=1	surprise	2.08 [−0.52, 4.95]	90.36	0.00 [−4.17, 2.86]	1.000
Qwen3-32B	Module M2	α=1	joy	−6.51 [−10.68, −2.60]	95.83	−14.58 [−23.44, −8.33]	0.024
Qwen3-32B	Module M3	α=1	anger	−3.65 [−6.25, −1.56]	94.27	−34.64 [−42.19, −27.34]	0.021
Qwen3-32B	Module M4	α=1	anger	−2.60 [−4.43, −0.78]	83.07	−14.06 [−19.01, −9.64]	0.088
Llama-3.3-70B	DiffMean	L=40, c=0.1	anger	1.30 [−0.78, 3.65]	96.61	0.00 [−3.65, 2.34]	1.000
Llama-3.3-70B	DiffMean	L=40, c=0.1	disgust	1.30 [−0.26, 3.39]	95.05	−1.04 [−5.21, 1.56]	1.000
Llama-3.3-70B	DiffMean	L=40, c=0.1	fear	−0.78 [−3.91, 2.34]	96.35	−3.13 [−9.11, 0.00]	1.000
Llama-3.3-70B	DiffMean	L=40, c=0.1	joy	4.95 [1.04, 8.85]	96.88	4.95 [−1.04, 8.33]	0.317
Llama-3.3-70B	DiffMean	L=40, c=0.1	sadness	0.26 [−2.34, 2.60]	95.83	−0.78 [−6.25, 1.30]	1.000
Llama-3.3-70B	DiffMean	L=40, c=0.1	surprise	0.26 [−1.82, 2.34]	95.83	−2.86 [−7.29, 0.00]	1.000
Llama-3.3-70B	Random	L=40, c=0.1	anger	2.60 [0.78, 4.95]	97.14	2.08 [−3.13, 4.43]	0.427
Llama-3.3-70B	Random	L=40, c=0.1	disgust	−0.26 [−1.30, 0.52]	96.35	−2.60 [−7.03, −1.04]	1.000
Llama-3.3-70B	Random	L=40, c=0.1	fear	−0.52 [−3.91, 2.60]	96.88	−2.08 [−8.33, 1.04]	1.000
Llama-3.3-70B	Random	L=40, c=0.1	joy	−0.52 [−4.17, 3.13]	96.61	−2.86 [−9.64, 0.78]	1.000
Llama-3.3-70B	Random	L=40, c=0.1	sadness	1.30 [−0.26, 3.13]	97.14	0.52 [−3.39, 1.82]	1.000
Llama-3.3-70B	Random	L=40, c=0.1	surprise	−0.26 [−1.82, 1.30]	95.57	−2.34 [−7.03, −0.78]	1.000
Llama-3.3-70B	Module M0	α=0.25	sadness	−0.78 [−3.39, 1.56]	95.31	−3.91 [−9.38, −0.52]	1.000
Llama-3.3-70B	Module M1	α=0.25	anger	−2.60 [−4.95, −0.78]	0.00	−45.83 [−53.91, −38.02]	0.264
Llama-3.3-70B	Module M2	α=0.25	sadness	−0.78 [−2.35, 0.52]	96.61	−1.56 [−5.73, −0.52]	1.000
Llama-3.3-70B	Module M3	α=0.25	anger	0.00 [−1.04, 1.04]	96.09	−0.78 [−4.69, 0.00]	1.000
Llama-3.3-70B	Module M4	α=0.25	anger	0.00 [−1.56, 1.56]	97.40	−3.39 [−8.33, −0.52]	1.000
Qwen2.5-72B	DiffMean	L=60, c=0.4	anger	4.69 [1.82, 8.07]	80.47	−0.52 [−5.47, 4.43]	0.032
Qwen2.5-72B	DiffMean	L=60, c=0.4	disgust	2.60 [0.26, 5.47]	95.05	−2.60 [−7.03, 1.04]	1.000
Qwen2.5-72B	DiffMean	L=60, c=0.4	fear	−1.04 [−4.43, 2.34]	92.19	−41.93 [−49.48, −34.38]	1.000
Qwen2.5-72B	DiffMean	L=60, c=0.4	joy	13.28 [7.29, 19.27]	84.64	3.65 [−4.17, 11.46]	8.50×10^{-4}
Qwen2.5-72B	DiffMean	L=60, c=0.4	sadness	1.04 [−1.04, 3.13]	96.88	−1.82 [−5.73, 1.04]	1.000
Qwen2.5-72B	DiffMean	L=60, c=0.4	surprise	16.41 [11.46, 21.61]	97.66	−3.91 [−11.98, 3.91]	8.50×10^{-4}
Qwen2.5-72B	Random	L=60, c=0.4	anger	0.26 [−0.52, 1.30]	96.88	−2.08 [−5.73, −0.52]	1.000
Qwen2.5-72B	Random	L=60, c=0.4	disgust	−1.04 [−3.13, 0.78]	98.44	−5.47 [−10.94, −2.60]	1.000
Qwen2.5-72B	Random	L=60, c=0.4	fear	1.30 [−1.30, 3.91]	95.83	−0.26 [−5.73, 2.60]	1.000
Qwen2.5-72B	Random	L=60, c=0.4	joy	6.25 [1.56, 10.94]	96.88	4.17 [−1.30, 8.85]	0.164
Qwen2.5-72B	Random	L=60, c=0.4	sadness	−1.56 [−3.91, 0.52]	98.44	−8.07 [−15.36, −2.08]	1.000
Qwen2.5-72B	Random	L=60, c=0.4	surprise	−0.52 [−2.61, 1.82]	97.92	−5.99 [−10.16, −2.08]	1.000
Qwen2.5-72B	Module M0	α=1	anger	0.26 [−0.52, 1.30]	94.79	−2.60 [−9.11, −0.52]	1.000
Qwen2.5-72B	Module M1	α=1	sadness	−0.78 [−2.60, 1.04]	94.27	−2.60 [−7.81, −0.78]	1.000
Qwen2.5-72B	Module M2	α=1	anger	0.78 [0.00, 1.82]	97.14	−1.56 [−6.51, 0.52]	1.000
Qwen2.5-72B	Module M3	α=1	anger	−0.78 [−2.08, 0.00]	100.00	−96.61 [−99.48, −93.49]	1.000
Qwen2.5-72B	Module M4	α=1	joy	3.65 [−0.78, 8.07]	96.35	1.04 [−5.21, 5.73]	1.000

Table 19: Matched evaluation in Qwen2.5-VL-7B and Qwen3-VL-32B, continuing Table 18; settings, units, and the Holm-17 family are the same.

Model	Method	Setting	Target	ΔSSR [95% CI]	Gate	Spec. [95% CI]	Holm-17 p
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	anger	0.52 [−1.04, 2.34]	95.05	−1.82 [−5.99, 0.52]	1.000
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	disgust	1.30 [0.00, 2.60]	96.61	0.26 [−3.13, 2.34]	1.000
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	fear	2.34 [−0.78, 5.47]	93.23	−0.26 [−4.43, 3.39]	1.000
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	joy	3.91 [−0.26, 8.33]	89.32	4.69 [−0.53, 8.85]	1.000
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	sadness	−0.78 [−3.13, 1.30]	95.05	−2.34 [−8.33, 0.00]	1.000
Qwen2.5-VL-7B	DiffMean	L=21, c=0.2	surprise	−0.26 [−1.82, 1.04]	88.54	−1.56 [−5.21, 0.26]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	anger	−0.26 [−1.82, 1.04]	97.14	−4.43 [−10.16, −0.78]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	disgust	1.30 [−0.26, 3.13]	97.66	−3.39 [−8.85, 1.04]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	fear	0.00 [−2.60, 2.60]	94.27	−4.95 [−12.24, 0.78]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	joy	1.30 [−2.60, 5.21]	97.40	0.00 [−8.59, 4.95]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	sadness	−0.78 [−2.86, 1.30]	92.71	−6.25 [−12.24, −1.56]	1.000
Qwen2.5-VL-7B	Random	L=21, c=0.2	surprise	−0.52 [−2.60, 1.30]	98.70	−9.11 [−14.84, −3.65]	1.000
Qwen2.5-VL-7B	Module M0	α=0.5	anger	0.52 [−0.78, 1.82]	92.71	−1.82 [−5.99, 0.00]	1.000
Qwen2.5-VL-7B	Module M1	α=0.5	anger	−1.30 [−3.39, 0.00]	66.41	−14.84 [−22.66, −7.55]	1.000
Qwen2.5-VL-7B	Module M2	α=0.5	sadness	−2.86 [−4.95, −1.04]	97.40	−5.47 [−11.20, −3.39]	0.065
Qwen2.5-VL-7B	Module M3	α=0.5	sadness	−2.08 [−4.17, −0.26]	86.20	−5.47 [−9.90, −2.86]	1.000
Qwen2.5-VL-7B	Module M4	α=0.5	anger	0.00 [−1.30, 1.30]	91.15	−1.82 [−4.69, −0.52]	1.000
Qwen3-VL-32B	DiffMean	L=48, c=0.4	anger	9.11 [5.47, 13.02]	68.49	5.99 [−1.56, 11.20]	8.50×10^{-4}
Qwen3-VL-32B	DiffMean	L=48, c=0.4	disgust	12.24 [8.07, 16.67]	89.06	1.04 [−5.99, 6.25]	8.50×10^{-4}
Qwen3-VL-32B	DiffMean	L=48, c=0.4	fear	8.85 [3.91, 13.80]	83.85	−2.86 [−11.20, 5.21]	0.0060
Qwen3-VL-32B	DiffMean	L=48, c=0.4	joy	57.81 [50.52, 64.58]	84.90	68.75 [61.72, 75.26]	8.50×10^{-4}
Qwen3-VL-32B	DiffMean	L=48, c=0.4	sadness	1.30 [−1.04, 3.91]	84.90	−0.52 [−5.47, 1.56]	1.000
Qwen3-VL-32B	DiffMean	L=48, c=0.4	surprise	66.41 [60.94, 71.88]	88.54	75.78 [69.79, 80.47]	8.50×10^{-4}
Qwen3-VL-32B	Random	L=48, c=0.4	anger	12.24 [7.55, 16.93]	93.23	7.81 [0.51, 13.80]	8.50×10^{-4}
Qwen3-VL-32B	Random	L=48, c=0.4	disgust	−1.56 [−3.13, −0.26]	85.94	−33.85 [−41.15, −26.82]	0.392
Qwen3-VL-32B	Random	L=48, c=0.4	fear	−7.55 [−10.94, −4.43]	39.84	−60.16 [−67.45, −53.13]	8.50×10^{-4}
Qwen3-VL-32B	Random	L=48, c=0.4	joy	−9.90 [−14.58, −5.99]	92.71	−41.93 [−51.82, −32.29]	8.50×10^{-4}
Qwen3-VL-32B	Random	L=48, c=0.4	sadness	26.04 [20.57, 31.25]	85.42	22.14 [13.80, 30.47]	8.50×10^{-4}
Qwen3-VL-32B	Random	L=48, c=0.4	surprise	−2.60 [−4.95, −0.78]	95.83	−35.42 [−43.23, −27.86]	0.116
Qwen3-VL-32B	Module M0	α=0.25	anger	1.04 [−0.52, 2.60]	84.38	−0.52 [−4.69, 0.78]	1.000
Qwen3-VL-32B	Module M1	α=0.25	joy	38.28 [31.51, 44.79]	100.00	10.94 [1.30, 20.31]	8.50×10^{-4}
Qwen3-VL-32B	Module M2	α=0.25	anger	0.52 [−0.52, 1.56]	84.11	−4.43 [−10.16, 0.00]	1.000
Qwen3-VL-32B	Module M3	α=0.25	sadness	−0.52 [−2.34, 1.30]	80.47	−2.08 [−6.77, −0.26]	1.000
Qwen3-VL-32B	Module M4	α=0.25	joy	0.00 [−2.08, 1.82]	82.81	0.00 [−3.39, 1.82]	1.000

Table 20: Cross-modal intervention test on 240 CREMA-D frames at each model's development-selected layer (7B: Qwen2.5-VL-7B, layer 21; 32B: Qwen3-VL-32B, layer 48). Δp is the change in the target-label probability and Spec. the margin over the largest off-target change, in percentage points with actor-bootstrap intervals; DiffMean−Random is the paired difference, with Holm-20 p over five targets and four strengths within each model. The last column is the change in the share of frames predicted as the target, in percentage points.

Model	Target	c	DiffMean Δp [95% CI]	Spec. [95% CI]	Random Δp	DiffMean−Random [95% CI]	Holm-20 p	Δ share
7B	anger	0.05	0.16 [0.06, 0.27]	−2.63 [−3.15, −2.14]	−0.45	0.60 [0.34, 0.92]	0.0010	0.00
7B	anger	0.1	0.36 [0.18, 0.58]	−4.15 [−5.07, −3.31]	−0.73	1.09 [0.61, 1.65]	0.0010	0.42
7B	anger	0.2	1.45 [0.90, 2.10]	−4.25 [−5.58, −2.97]	−1.77	3.22 [2.00, 4.62]	0.0010	0.42
7B	anger	0.4	−0.14 [−0.40, 0.08]	−3.93 [−4.78, −3.13]	−4.96	4.81 [2.94, 6.95]	0.0010	−0.42
7B	disgust	0.05	0.18 [0.10, 0.27]	−2.08 [−2.46, −1.71]	0.72	−0.54 [−0.71, −0.39]	0.0010	0.00
7B	disgust	0.1	0.43 [0.25, 0.64]	−3.35 [−4.02, −2.72]	1.65	−1.23 [−1.60, −0.90]	0.0010	0.00
7B	disgust	0.2	0.99 [0.64, 1.39]	−3.95 [−4.85, −3.08]	3.80	−2.81 [−3.61, −2.11]	0.0010	0.42
7B	disgust	0.4	4.08 [3.05, 5.21]	−2.43 [−3.45, −1.40]	5.41	−1.34 [−2.06, −0.55]	0.014	3.33
7B	fear	0.05	−0.04 [−0.10, −0.01]	−3.37 [−3.93, −2.84]	−0.05	0.00 [−0.05, 0.05]	1.000	0.00
7B	fear	0.1	−0.09 [−0.20, −0.02]	−5.95 [−6.97, −5.00]	−0.07	−0.02 [−0.12, 0.07]	1.000	0.00
7B	fear	0.2	−0.13 [−0.31, −0.02]	−9.72 [−11.20, −8.33]	−0.14	0.01 [−0.12, 0.12]	1.000	0.00
7B	fear	0.4	0.45 [0.33, 0.59]	−10.96 [−12.57, −9.41]	0.42	0.03 [−0.20, 0.24]	1.000	0.00
7B	joy	0.05	4.15 [3.57, 4.74]	4.04 [3.45, 4.64]	1.55	2.60 [2.10, 3.13]	0.0010	2.08
7B	joy	0.1	11.19 [9.81, 12.62]	11.15 [9.76, 12.60]	2.95	8.24 [7.00, 9.53]	0.0010	9.17
7B	joy	0.2	30.07 [27.19, 32.96]	30.10 [27.21, 32.99]	−0.51	30.59 [27.59, 33.57]	0.0010	35.00
7B	joy	0.4	11.47 [8.95, 14.01]	8.89 [6.49, 11.31]	−14.48	25.95 [22.57, 29.39]	0.0010	14.58
7B	sadness	0.05	−0.86 [−1.16, −0.60]	−3.65 [−4.35, −3.00]	0.84	−1.70 [−2.22, −1.23]	0.0010	−0.42
7B	sadness	0.1	−1.83 [−2.43, −1.29]	−6.60 [−7.94, −5.37]	0.49	−2.32 [−3.08, −1.63]	0.0010	−2.08
7B	sadness	0.2	−2.82 [−3.73, −2.00]	−9.59 [−11.56, −7.79]	−2.38	−0.44 [−0.81, −0.07]	0.069	−3.75
7B	sadness	0.4	−1.99 [−2.67, −1.35]	−8.96 [−10.32, −7.59]	3.34	−5.33 [−6.09, −4.61]	0.0010	−2.50
32B	anger	0.05	1.28 [0.90, 1.69]	−0.75 [−1.43, −0.08]	1.22	0.06 [−0.54, 0.67]	1.000	0.83
32B	anger	0.1	2.17 [1.54, 2.84]	−1.75 [−3.00, −0.52]	1.73	0.44 [−0.71, 1.64]	1.000	0.83
32B	anger	0.2	6.17 [4.66, 7.83]	1.95 [−0.06, 4.09]	−3.29	9.47 [7.02, 12.21]	0.0010	5.42
32B	anger	0.4	17.54 [14.42, 20.76]	15.82 [12.58, 19.14]	−14.56	32.10 [26.77, 37.61]	0.0010	16.67
32B	disgust	0.05	2.26 [1.54, 3.05]	0.80 [−0.02, 1.70]	0.17	2.09 [1.38, 2.90]	0.0010	2.50
32B	disgust	0.1	4.65 [3.31, 6.10]	1.93 [0.34, 3.62]	−0.44	5.08 [3.54, 6.80]	0.0010	3.33
32B	disgust	0.2	4.01 [2.51, 5.61]	−0.91 [−3.07, 1.26]	−2.78	6.79 [4.49, 9.33]	0.0010	2.92
32B	disgust	0.4	8.00 [5.80, 10.30]	0.25 [−2.22, 2.62]	−8.49	16.49 [13.18, 19.87]	0.0010	6.25
32B	fear	0.05	2.26 [1.54, 3.11]	1.02 [0.17, 1.98]	−0.78	3.04 [2.05, 4.18]	0.0010	1.67
32B	fear	0.1	5.22 [3.67, 7.01]	3.10 [1.35, 5.09]	−1.97	7.19 [5.20, 9.43]	0.0010	4.17
32B	fear	0.2	9.42 [6.96, 12.20]	6.19 [3.32, 9.30]	−6.90	16.33 [12.81, 20.21]	0.0010	9.17
32B	fear	0.4	2.53 [1.33, 3.89]	−8.62 [−11.29, −6.06]	−16.15	18.68 [14.12, 23.33]	0.0010	1.25
32B	joy	0.05	0.73 [0.30, 1.25]	−1.38 [−2.12, −0.63]	1.50	−0.77 [−1.66, 0.06]	0.589	1.25
32B	joy	0.1	1.35 [0.64, 2.18]	−2.66 [−3.93, −1.41]	3.02	−1.67 [−3.48, 0.03]	0.589	1.67
32B	joy	0.2	4.74 [3.17, 6.45]	−2.29 [−4.70, 0.09]	3.86	0.88 [−1.41, 3.12]	1.000	3.75
32B	joy	0.4	36.71 [32.01, 41.41]	32.44 [27.45, 37.52]	−10.13	46.84 [42.48, 51.21]	0.0010	39.17
32B	sadness	0.05	−0.46 [−0.82, −0.16]	−2.41 [−3.12, −1.75]	1.70	−2.16 [−3.01, −1.45]	0.0010	0.00
32B	sadness	0.1	−0.62 [−1.25, −0.05]	−4.42 [−5.78, −3.17]	4.42	−5.04 [−6.79, −3.54]	0.0010	−0.83
32B	sadness	0.2	−0.49 [−1.36, 0.32]	−6.11 [−8.10, −4.33]	7.30	−7.78 [−9.96, −5.67]	0.0010	−1.67
32B	sadness	0.4	−0.89 [−1.79, −0.03]	−8.61 [−10.64, −6.64]	−8.76	7.87 [4.73, 11.22]	0.0010	−0.83

Table 21: Each vision-language model's own modules applied during CREMA-D frame classification. Δp and Spec. are in percentage points with actor-bootstrap intervals, and Δ share is the change in the share of frames predicted as the target, in percentage points; accuracy is in percent, with the unsteered accuracy in parentheses. Holm-15 adjusts the 15 module conditions within each model.

Model	Mod.	Target	α	Δp [95% CI]	Spec. [95% CI]	Holm-15 p	Δ share	Acc. (base)
Qwen2.5-VL-7B	M0	anger	0.5	1.41 [0.95, 1.93]	0.33 [−0.18, 0.91]	7.50×10^{-4}	0.00	32.50 (32.92)
Qwen2.5-VL-7B	M0	anger	1	2.65 [1.86, 3.54]	−0.60 [−1.62, 0.55]	7.50×10^{-4}	0.83	32.50 (32.92)
Qwen2.5-VL-7B	M0	anger	2	10.35 [8.83, 11.93]	−7.03 [−8.92, −5.09]	7.50×10^{-4}	7.50	32.92 (32.92)
Qwen2.5-VL-7B	M1	anger	0.5	−4.74 [−6.90, −2.87]	−23.53 [−27.50, −19.70]	7.50×10^{-4}	−7.92	19.17 (32.92)
Qwen2.5-VL-7B	M1	anger	1	−4.26 [−6.93, −1.91]	−26.35 [−29.40, −23.64]	0.0044	−7.92	16.67 (32.92)
Qwen2.5-VL-7B	M1	anger	2	−4.99 [−7.67, −2.64]	−36.46 [−39.15, −34.10]	8.00×10^{-4}	−7.92	17.08 (32.92)
Qwen2.5-VL-7B	M2	sadness	0.5	−1.48 [−2.00, −0.99]	−4.84 [−5.82, −3.89]	7.50×10^{-4}	−2.08	32.50 (32.92)
Qwen2.5-VL-7B	M2	sadness	1	14.66 [13.38, 15.98]	−19.89 [−21.92, −17.72]	7.50×10^{-4}	7.50	31.67 (32.92)
Qwen2.5-VL-7B	M2	sadness	2	7.61 [4.72, 10.30]	−45.43 [−49.49, −41.51]	7.50×10^{-4}	−7.08	16.67 (32.92)
Qwen2.5-VL-7B	M3	sadness	0.5	2.63 [2.15, 3.13]	0.88 [0.25, 1.51]	7.50×10^{-4}	1.67	33.75 (32.92)
Qwen2.5-VL-7B	M3	sadness	1	6.08 [5.08, 7.14]	1.65 [0.21, 3.11]	7.50×10^{-4}	4.17	33.33 (32.92)
Qwen2.5-VL-7B	M3	sadness	2	5.96 [3.78, 8.01]	−14.13 [−16.84, −11.59]	7.50×10^{-4}	−3.75	38.33 (32.92)
Qwen2.5-VL-7B	M4	anger	0.5	0.17 [0.03, 0.32]	−1.55 [−1.91, −1.18]	0.046	0.00	33.75 (32.92)
Qwen2.5-VL-7B	M4	anger	1	1.18 [0.77, 1.64]	−3.28 [−4.14, −2.44]	7.50×10^{-4}	0.00	34.17 (32.92)
Qwen2.5-VL-7B	M4	anger	2	2.62 [0.07, 4.87]	−66.95 [−69.57, −64.46]	0.038	−7.92	16.67 (32.92)
Qwen3-VL-32B	M0	anger	0.5	0.07 [−0.61, 0.74]	−3.92 [−5.45, −2.52]	0.586	0.00	54.17 (57.08)
Qwen3-VL-32B	M0	anger	1	0.54 [−0.99, 2.06]	−8.18 [−11.32, −5.23]	0.354	1.25	52.92 (57.08)
Qwen3-VL-32B	M0	anger	2	−3.18 [−6.07, −0.53]	−24.50 [−30.58, −18.68]	0.277	−2.08	54.17 (57.08)
Qwen3-VL-32B	M1	joy	0.5	23.17 [19.41, 26.71]	2.24 [−1.61, 5.93]	7.50×10^{-4}	86.67	16.67 (57.08)
Qwen3-VL-32B	M1	joy	1	29.54 [25.72, 33.15]	8.41 [4.49, 12.16]	7.50×10^{-4}	86.67	16.67 (57.08)
Qwen3-VL-32B	M1	joy	2	23.67 [19.84, 27.27]	0.51 [−3.39, 4.27]	7.50×10^{-4}	86.67	16.67 (57.08)
Qwen3-VL-32B	M2	anger	0.5	2.30 [1.10, 3.57]	−4.02 [−6.44, −1.79]	0.086	2.50	53.75 (57.08)
Qwen3-VL-32B	M2	anger	1	25.32 [21.67, 28.78]	6.55 [2.99, 9.91]	7.50×10^{-4}	70.83	25.83 (57.08)
Qwen3-VL-32B	M2	anger	2	4.59 [−0.07, 8.85]	−34.13 [−39.34, −29.24]	0.188	−18.75	16.67 (57.08)
Qwen3-VL-32B	M3	sadness	0.5	0.79 [0.34, 1.32]	−2.57 [−3.66, −1.56]	0.052	0.83	54.17 (57.08)
Qwen3-VL-32B	M3	sadness	1	1.23 [0.26, 2.38]	−6.01 [−8.23, −3.85]	0.277	0.00	55.00 (57.08)
Qwen3-VL-32B	M3	sadness	2	4.04 [1.85, 6.48]	−13.31 [−17.27, −9.36]	0.011	4.17	52.08 (57.08)
Qwen3-VL-32B	M4	joy	0.5	−0.15 [−0.29, −0.03]	−2.38 [−3.00, −1.82]	0.079	0.00	57.08 (57.08)
Qwen3-VL-32B	M4	joy	1	−0.27 [−0.51, −0.08]	−4.50 [−5.74, −3.41]	0.052	0.00	56.25 (57.08)
Qwen3-VL-32B	M4	joy	2	−0.44 [−0.75, −0.19]	−9.19 [−11.67, −7.03]	0.0017	0.00	57.08 (57.08)

E VALIDATING OUTPUT QUALITY

This appendix checks whether classifier-judged gains reflect readable text, using a rule-based screen for degenerate output on every generation and an AI-assisted readability review of a stratified sample. Three results support the main analyses. First, unsteered outputs are clean: the screen flags none of the 1,920 baseline and 160 calibration outputs of the five matched evaluations, and the review rates all 135 sampled baselines mostly readable. Second, the mean-difference gains rest mostly on unflagged text: the screen flags 195 of the 11,520 mean-difference outputs (1.69%), against 1,029 of the 11,520 random-direction outputs (8.93%); 8 of the 12 mean-difference gains that pass Holm-17 contain no flagged hit, and 1,372 of their 1,535 target hits (89.38%) are unflagged, with the largest flagged share in Qwen3-VL-32B surprise (72 of 265). Third, the screen and review locate where a likelihood gate alone is insufficient, in random directions and some modules, which is why Section 6.4 reports quality checks beyond likelihood.

Screening the matched evaluation. The screen, applied to the 6,912 matched-evaluation outputs of each model (128 prompts $\times$ 3 seeds $\times$ 18 conditions), flags no unsteered baseline or calibration output in any model and no output of Llama-3.3-70B (Table 22). Flags concentrate in random directions and a few modules: Qwen2.5-72B module 3 is flagged on all 384 outputs, and Qwen3-VL-32B's random anger and fear directions on 260 and 210. Table 23 decomposes every positive gain that passes Holm-17 into the gate-passing target-labeled outputs of the steered and baseline runs, so the flagged share of each gain can be read directly. The rules target repetition and symbol runs, so their counts are a lower bound for other degenerate forms: they flag 48 of the 185 module-1 joy outputs and 14 of the 113 random-sadness outputs of Qwen3-VL-32B, whereas every one of 35 outputs we read from each of these conditions was a string of isolated characters or of repeated "or", asterisks, and digits.

Table 22: Screen flags on the matched-evaluation outputs: flagged/outputs, with flagged gate passes in parentheses. DiffMean and random cells pool six targets, module cells five modules. The 32 calibration outputs per model are also unflagged.

Model	Baseline	DiffMean	Random	Modules
Qwen3-32B	0/384 (0)	101/2,304 (101)	430/2,304 (416)	94/1,920 (89)
Llama-3.3-70B	0/384 (0)	0/2,304 (0)	0/2,304 (0)	0/1,920 (0)
Qwen2.5-72B	0/384 (0)	0/2,304 (0)	0/2,304 (0)	384/1,920 (384)
Qwen2.5-VL-7B	0/384 (0)	0/2,304 (0)	0/2,304 (0)	188/1,920 (166)
Qwen3-VL-32B	0/384 (0)	94/2,304 (93)	599/2,304 (447)	73/1,920 (73)

Table 23: Every positive gain that passes Holm-17 in the matched evaluation (Tables 18 and 19). Hits are gate-passing outputs the classifier labels with the target, out of 384; ΔSSR = (steered $-$ baseline hits)$/384$ in percentage points. Flagged hits are steered hits that trigger any screen rule.

Model	Condition	Target	ΔSSR	Steered hits	Baseline hits	Flagged hits
Qwen3-32B	DiffMean	anger	10.68	55	14	0
Qwen3-32B	DiffMean	disgust	17.19	70	4	0
Qwen3-32B	DiffMean	joy	55.73	254	40	31
Qwen3-32B	DiffMean	surprise	69.27	273	7	46
Qwen2.5-72B	DiffMean	anger	4.69	21	3	0
Qwen2.5-72B	DiffMean	joy	13.28	103	52	0
Qwen2.5-72B	DiffMean	surprise	16.41	72	9	0
Qwen3-VL-32B	DiffMean	anger	9.11	44	9	0
Qwen3-VL-32B	DiffMean	disgust	12.24	53	6	0
Qwen3-VL-32B	DiffMean	fear	8.85	65	31	0
Qwen3-VL-32B	DiffMean	joy	57.81	260	38	14
Qwen3-VL-32B	DiffMean	surprise	66.41	265	10	72
Qwen3-VL-32B	Module M1	joy	38.28	185	38	48
Qwen3-VL-32B	Random	anger	12.24	56	9	47
Qwen3-VL-32B	Random	sadness	26.04	113	13	14

Screening the historical module outputs. The simple diagnostic flags empty text, nonempty text without alphanumeric characters, or a single word occupying at least 80% of at least ten word tokens. The expanded screen additionally flags a generated token ID occupying at least 80% of at least 20 token IDs, at least ten consecutive identical words, unique word-bigram fraction at most 20% in at least 20 words, or at least 90% nonalphanumeric characters among at least 20 nonspace characters. Exact thresholds and Unicode handling are preserved in the analysis code. The rules apply uniformly to baseline and intervention outputs and detect degeneracy rather than certify coherence. All 1,152 baseline outputs pass the screens, and 1,108 pass the original likelihood gate. Among the 17,280 intervention outputs, 2,641 trigger the simple screen and 3,794 trigger any screen (21.96%); of the 9,502 gate passes, 2,482 and 3,318 trigger the respective screens.

AI-assisted review: sampling and masking. Before inspecting sampled text or gate status for rating, the review procedure selects three paired draws per model/module/strength, covering seeds 0, 1, and 2, using sampling seed 20260922. The sample contains 45 distinct prompts per model, with each assigned to one module–strength–seed setting, giving 135 pairs and 270 displayed outputs. No prompt is reused across settings within a model in this review sample. Baseline and intervention outputs are shuffled under opaque identifiers; the reviewer sees the prompt and saved text without condition metadata, and ratings are frozen before the condition mapping is joined. The AI reviewer had already seen aggregate results and some failure examples, and style can reveal intervention status, so we describe the procedure as metadata masking. LLM annotations can approach human-written quality (Huang et al., 2023), and a blinded human rater checks this review (Appendix G), matching 11 of its 12 shared readability ratings.

Table 24: Model breakdown of the metadata-masked *AI-assisted* review of 135 paired module cases. Readable means rubric score 2; the final column counts steered texts scored below 2 among gate passes. The AI reviewer had prior aggregate knowledge; Appendix G reports the blinded human check.

Model	Baseline readable	Steered readable	Steered gate pass	Failed readability / pass
Qwen3-32B	45/45	9/45	32/45	24/32
Llama-3.3-70B	45/45	8/45	14/45	6/14
Qwen2.5-72B	45/45	6/45	29/45	23/29

Rubric and ratings. The scale is coarse because LLM quality ratings align with human ratings better on coarse than on fine-grained scales (Huang et al., 2024). Readability 2 means mostly coherent, interpretable text; 1 means partly intelligible but materially malformed; 0 means unreadable or nonlinguistic. Coherent non-English text is not automatically scored as unreadable. Continuation compliance is separately marked yes, no, or unclear, and requires a relevant continuation rather than a role marker, unrelated response, or bare ending token. No emotion labels are newly assigned. All 135 baselines are rated mostly readable. The 135 historical module interventions, sampled at $\alpha \in \{2, 4, 8\}$, receive 101 zeros, 11 ones, and 23 twos; of the 23 readable outputs, 18 comply, two have unclear compliance, and three do not comply. The screen and the review complement each other: the screen flags none of the 23 readable outputs and catches 25 of the 112 outputs scored below 2.

Table 24 breaks the ratings down by model; with three examples per setting, the review is a qualitative diagnostic. Here and below, Qwen3, Llama, and Qwen2.5 denote Qwen3-32B, Llama-3.3-70B, and Qwen2.5-72B. The paired readable-rate differences are -80.00 points for Qwen3 (nominal interval $[-91.11, -66.67]$), -82.22 for Llama ($[-93.33, -71.11]$), and -86.67 for Qwen2.5 ($[-95.56, -75.56]$). These model-level intervals describe the fixed, equally weighted module/strength mixture rather than each setting. Six readable baselines fail the original gate, so the likelihood proxy also rejects acceptable outputs.

Illustrative gate failures. Examples selected after rating show the failure types. A Qwen3 module-2 continuation at $\alpha = 4$ consists almost entirely of commas and a digit; it passes the original gate with mean negative log likelihood 3.526 but is rated unreadable. A Qwen2.5 module-3 continuation at $\alpha = 2$ repeats a short Cyrillic fragment, passes with mean negative log likelihood 0.332, and is also rated unreadable; the same module yields the same classifier outcome at every tested strength, down to $\alpha = 0.25$ (gate-pass rate 100%, specificity margin -96.61), and at $\alpha = 2$, 4, and 8 all 384

of its outputs trigger the repetition screen. A Llama module-3 output at $\alpha = 4$ begins readably but degenerates into a malformed identity statement; it passes with mean negative log likelihood 2.920 and evades the screen, receiving readability 1. In the Qwen3-VL-32B matched evaluation, a module-1 continuation that the classifier counts toward the joy gain reads "y _ q _ y y q i *" (spacing collapsed), and the random-sadness continuations repeat "or", asterisks, and digits; all module-1 outputs pass the gate. The rates above come from the sampled and full populations; these cases illustrate the failure types.

F Matched Random-Ablation Pilot

The pilot evaluates the 301 held-out classification items for each text model. Test labels comprise 60 anger, 23 disgust, 35 fear, 45 joy, 59 neutral, 45 sadness, and 34 surprise. Zero ablation is applied at the feed-forward down-projection input, and each control samples the same number of units in each layer as the corresponding learned module. There are five random draws per module, 75 control conditions in total, and 27,993 item-condition predictions across all baselines and interventions; this count does not represent independent examples.

Let $d_e = a_e^{\text{base}} - a_e^{\text{ablated}}$ be the signed class-accuracy drop. We report signed target drops and define

$$S_{\max} = \frac{\max(0, d_e)}{\max_{j \neq e} \max(0, d_j) + 10^{-6}}. \tag{5}$$

This maximum-off-target denominator is stricter than averaging off-target drops. Negative signed drops are retained in Table 25, although the ratio clips them to zero. Two Qwen2.5 module-2 random draws are epsilon-dominated rather than evidence of extraordinary selectivity. The raw add-one random-control p compares each learned ratio with the five matched random ratios, so its minimum is $1/6$ and no learned module can pass a 0.05 test. Figure 8 compares every learned ablation with its random controls. Only Qwen3 module 1 (surprise) and module 2 (joy) have positive target drops larger than every off-target drop (38.24 versus 13.04 points, and 22.22 versus 20.59).

Several ablations instead improve target-class accuracy while harming other classes, and Table 26 shows why: they shift the output label prior. Ablating Qwen3 module 3 raises the share of anger predictions by 79.07 points and Llama module 1 by 85.71 points, and Qwen2.5 module 3 moves the predicted-label distribution by a total-variation distance of 97.67%. Matched random ablations produce shifts at least as large for Llama module 1 (98.01%) and Qwen2.5 module 3 (97.67%), and only 5 of the 15 learned modules exceed all their random controls, so these prior shifts are not specific to the learned modules. A class-accuracy drop metric cannot distinguish such prior shifts from selective damage. Baseline overall accuracies on the 301 items are 38.21%, 36.21%, and 34.22% for Qwen3, Llama, and Qwen2.5, respectively. The pilot therefore does not validate selective emotion modules.

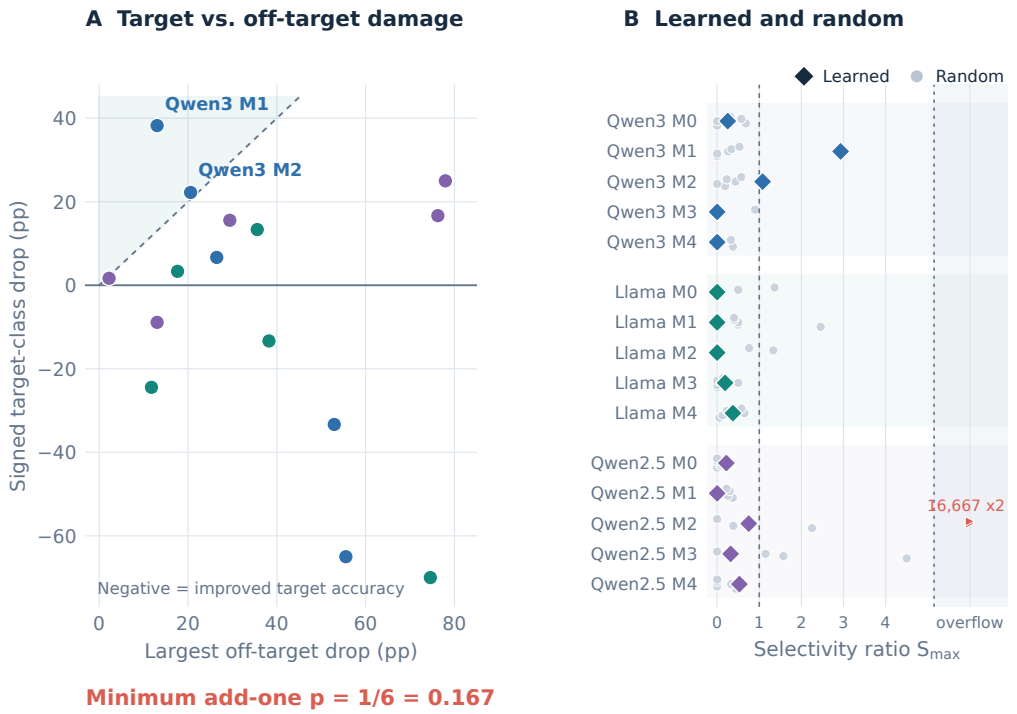

Figure 8: All 15 learned-module ablations and 75 matched random controls on 301 held-out items. (A) Signed target-class drop versus largest clipped off-target drop; above the equality line, target damage is greater. Negative signed drops improve target accuracy. (B) Learned $S_{\max}$ ratios (diamonds) and all five random ratios per module (gray points, vertically offset for visibility). Two Qwen2.5 M2 random ratios near 16,667 are epsilon-dominated and shown in a labeled overflow lane. Five draws imply a minimum add-one $p = 1/6$.

Table 25: All 15 learned-module ablations. Signed target drops and largest clipped off-target drops are percentage points; ratios are dimensionless. The minimum attainable raw p is $1/6$. Positive ratios alone do not demonstrate selective causal modules.

Model	Mod.	Target	Target drop	Max off-drop	$S_{\max}$	Raw p
Qwen3-32B	0	anger	6.67	26.47	0.252	0.667
Qwen3-32B	1	surprise	38.24	13.04	2.931	0.167
Qwen3-32B	2	joy	22.22	20.59	1.079	0.167
Qwen3-32B	3	anger	−65.00	55.56	0.000	1.000
Qwen3-32B	4	anger	−33.33	52.94	0.000	1.000
Llama-3.3-70B	0	sadness	−13.33	38.24	0.000	1.000
Llama-3.3-70B	1	anger	−70.00	74.58	0.000	1.000
Llama-3.3-70B	2	sadness	−24.44	11.76	0.000	1.000
Llama-3.3-70B	3	anger	3.33	17.65	0.189	0.333
Llama-3.3-70B	4	anger	13.33	35.59	0.375	0.500
Qwen2.5-72B	0	anger	16.67	76.27	0.219	0.167
Qwen2.5-72B	1	sadness	−8.89	13.04	0.000	1.000
Qwen2.5-72B	2	anger	1.67	2.22	0.750	0.667
Qwen2.5-72B	3	anger	25.00	77.97	0.321	0.833
Qwen2.5-72B	4	joy	15.56	29.41	0.529	0.167

Table 26: Label-prior shifts under ablation: change in the share of predictions equal to the module's target label (percentage points), total-variation (TV) distance between predicted-label distributions before and after ablation, and the largest distance among the five matched random controls (%). "Exceeds random" requires a strictly larger distance, so ties with a random control count as no.

Model	Mod.	Target	Δ target share (pp)	TV (%)	Max random TV (%)	Exceeds random
Qwen3-32B	M0	anger	−1.99	22.26	67.77	no
Qwen3-32B	M1	surprise	−11.30	17.61	18.94	no
Qwen3-32B	M2	joy	−8.64	22.26	28.24	no
Qwen3-32B	M3	anger	79.07	79.07	26.25	yes
Qwen3-32B	M4	anger	37.21	43.85	55.15	no
Llama-3.3-70B	M0	sadness	2.99	22.92	16.94	yes
Llama-3.3-70B	M1	anger	85.71	85.71	98.01	no
Llama-3.3-70B	M2	sadness	10.63	10.63	92.69	no
Llama-3.3-70B	M3	anger	−2.33	14.62	14.62	no
Llama-3.3-70B	M4	anger	−0.66	25.25	58.47	no
Qwen2.5-72B	M0	anger	−5.98	83.72	56.81	yes
Qwen2.5-72B	M1	sadness	9.30	11.30	97.67	no
Qwen2.5-72B	M2	anger	−1.66	3.65	3.32	yes
Qwen2.5-72B	M3	anger	−8.31	97.67	97.67	no
Qwen2.5-72B	M4	joy	−4.98	28.90	11.30	yes

G Single-Rater Human Check

One independent rater rated continuations in a self-contained offline booklet that showed only the original text and the continuation under opaque identifiers; model, method, target, gate status, and classifier label were hidden. The 74 items comprise 54 outputs sampled, regardless of gate or classifier label, from the Qwen3-32B, Qwen2.5-72B, and Qwen2.5-VL-7B matched-evaluation runs, which were complete when the booklet was built (Table 27); 12 outputs from the AI-assisted review (four rated readable, four partly readable, and four unreadable by the AI); and 8 of the 54 shown a second time without notice. For each item the rater gave readability on the Appendix E rubric, whether the text is a relevant continuation, and the emotion expressed most strongly among the classifier's seven labels, "cannot tell", or "other" with a required comment. The rater finished in 36.30 minutes (median 16 seconds of wall-clock time per item), reported no prior exposure, and changed no answer after first completing it.

The rater's readability matched the AI-assisted rating on 11 of the 12 shared outputs (Cohen's $\kappa = 0.875$; the one difference was 1 against the AI's 0), and all 8 repeated items received identical answers. All 30 DiffMean continuations (12 Qwen3-32B, 12 Qwen2.5-72B, and 6 Qwen2.5-VL-7B) were rated readable and 29 relevant, whereas all three Qwen3-32B random-direction continuations were rated unreadable or partly readable although most random-direction outputs pass the likelihood gate. The rater named the steering target for 2 of the 30 DiffMean items and perceived any emotion in 6, against none of the 10 baselines; the two Qwen3-32B joy items were marked "other" with the comments "It is like excitement." and "It is excitement." Agreement between the rater's emotion and the classifier label over the 54 session-1 outputs is weak (Cohen's $\kappa = 0.228$). In a second session, the same rater rated 44 outputs from the Llama-3.3-70B and Qwen3-VL-32B matched-evaluation runs, which finished after the first booklet was built, sampled in the same way (per model, 2 DiffMean outputs per target, 3 random-direction, 3 module, and 4 baseline outputs; Qwen3-VL-32B's random anger and sadness directions and module 1 were always included). No original text from the first session was reused, the booklet stated that it was a second session, and the rater finished in 20.58 minutes (median 12 seconds per item), reported prior exposure for 1 item, and changed no answer. All 24 DiffMean continuations were rated readable and relevant; the rater named the target emotion in none and perceived any emotion in 7, and marked one joy and one disgust continuation from Qwen3-VL-32B "other" with the comment "it is excitement." All three Qwen3-VL-32B random-direction outputs (anger, fear, and sadness) and the module-1 outputs of both models were rated unreadable, consistent with the screen in Appendix E. Agreement between the rater's emotion and the classifier label over these 44 outputs is $\kappa = 0.081$. With one rater and a few items per condition, these counts are a qualitative check rather than rate estimates.

Table 27: Single blinded human rater on 98 outputs sampled from the matched-evaluation runs of five models, in two sessions (session 2 added Llama-3.3-70B and Qwen3-VL-32B). Counts of items rated readable (score 2), relevant, naming the steering target, naming any emotion (a label other than neutral or "cannot tell"), and matching the classifier label, including matches on neutral. Random-direction and module targets are the target the condition was evaluated for; Qwen2.5-VL-7B had no random-direction items in this sample.

Session	Model	Condition	n	Readable	Relevant	Target named	Any emotion	Agrees with classifier
1	Qwen3-32B	DiffMean	12	12	11	0	3	4
1	Qwen3-32B	Random	3	0	0	0	0	0
1	Qwen3-32B	Module	3	3	3	1	1	1
1	Qwen3-32B	Baseline	4	4	4	–	0	3
1	Qwen2.5-72B	DiffMean	12	12	12	2	3	8
1	Qwen2.5-72B	Random	3	3	3	0	0	2
1	Qwen2.5-72B	Module	3	2	2	1	1	2
1	Qwen2.5-72B	Baseline	4	4	4	–	0	3
1	Qwen2.5-VL-7B	DiffMean	6	6	6	0	0	5
1	Qwen2.5-VL-7B	Module	2	2	2	0	0	2
1	Qwen2.5-VL-7B	Baseline	2	2	2	–	0	2
2	Llama-3.3-70B	DiffMean	12	12	12	0	2	8
2	Llama-3.3-70B	Random	3	3	3	0	0	3
2	Llama-3.3-70B	Module	3	2	2	0	0	1
2	Llama-3.3-70B	Baseline	4	4	4	–	0	3
2	Qwen3-VL-32B	DiffMean	12	12	12	0	5	4
2	Qwen3-VL-32B	Random	3	0	0	0	0	0
2	Qwen3-VL-32B	Module	3	2	2	0	0	2
2	Qwen3-VL-32B	Baseline	4	4	4	–	0	1

H Historical Analyses and Evidence Provenance

The behavioral and 128-prompt intervention records analyzed here supersede aggregates from earlier exploratory runs, which used different frames, smaller prompt sets, and a same-model judge with a heuristic quality check; those runs are not pooled with the records analyzed here. Earlier module archives also lack item identifiers, preventing the provenance filtering and group splits used here. Their exploratory outcomes remain in the archive rather than being presented as additional replications.

For completeness, an earlier standardized image-to-text ridge remapping on a separate 200-item, 384D signature archive increased matched-minus-mismatched concordance from -0.000035 to 0.007803 (199 fitted-pair shuffles, raw $p = 0.005$, Holm $p = 0.010$ across two remapping tests). A separate 60-item remapping gave -0.000513 with raw $p = 0.050$. These archives mix input provenance and do not permit group-aware identification checks. The positive geometric result is retained as exploratory evidence, but it is not an emotion-label transfer score or a substitute for the CREMA-D result. Whole-layer similarity indices such as centered kernel alignment (CKA) (Kornblith et al., 2019) can miss functionally relevant differences (Ding et al., 2021), so similarity alone would not establish a shared emotion circuit.

Source inventories and CPU-only build scripts verify the reported counts, paired tests, permutation tails, correction families, steering effects, review ratings, and ablation ratios. Generated tables and figures preserve the distinction between historical records, added experiments, and post-hoc CPU analyses.