# DO EMOTION CONCEPTS GENERALIZE ACROSS SOURCES, MODALITIES, AND ARCHITECTURES IN VISION-LANGUAGE MODELS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent studies suggest that large language models encode emotion concepts as structured internal representations, but most existing work focuses on text and a single architecture. Therefore, we ask, do emotion concepts generalize across sources, modalities, and architectures in vision–language models (VLMs)? To address this, we construct CMES (Cross-Modal Emotion Stimuli), a multi-source collection of emotion-conditioned stories, real facial expressions, synthetic portraits, and synthetic emotion-evoking scenes. For each stimulus source, we extract a separate set of six Ekman emotion vectors from each of three VLMs. We report four main findings as follows: 1) Image-derived emotion vectors form a low-dimensional geometry similar to that of text-derived vectors. Valence is relatively stable across sources, while arousal varies more. 2) Text- and image-derived emotion vectors have modest cosine similarity but still show held-out cross-modal correspondence. Text-derived vectors can also steer image interpretation. 3) Cross-architecture correspondence remains even when native cosine is near zero. Transformations estimated from generic ImageNet activations recover both correspondence and causal transfer without using the six emotion vectors or their labels. 4) After aligning representations across architectures, we construct a shared emotion subspace that preserves affective geometry and selective steering effects. The corresponding consensus emotion vectors also generalize to a held-out fourth architecture at two model sizes. These results suggest that emotion representations can share relational structure and causal effects across sources, modalities, and architectures, even when individual vector directions differ.

## 1 INTRODUCTION

As large language models (LLMs) develop, more studies are focusing on emotion understanding (Cheng et al., 2024; Lian et al., 2025; Xing et al., 2026). Meanwhile, recent interpretability work suggests that LLMs encode emotion concepts as structured internal representations. Using text-based stimuli, prior work has identified emotion vectors whose low-dimensional organization resembles human affective structure (Sofroniew et al., 2026; Reichman et al., 2026; Wang et al., 2026; van der Ben et al., 2026). Individual neurons and attention heads also carry emotion-related information (Wang et al., 2026; Zheng et al., 2026). In addition, interventions along emotion vectors can change emotional judgments and generated behavior (Zou et al., 2025; Sofroniew et al., 2026; Reichman et al., 2026). Recent work also traces emotion-related information flow within individual VLMs (Zhang et al., 2026). However, most existing evidence is based on text and a single model. It remains unclear how emotion representations change across stimulus sources and modalities, and whether their geometry and causal effects are preserved across architectures.

Researchers have also studied whether steering vectors can be transferred across models. Activation steering can modify refusal behavior and instruction following within individual models (Arditi et al., 2024; Lee et al., 2025; Stolfo et al., 2025). More recent studies transfer steering vectors for concepts and behavioral attributes across independently trained text LLMs through learned transformations (Huang et al., 2025; Chen et al., 2025; Agarwal, 2026). However, these results leave open how related emotion concepts are organized as a group across text and image sources, and whether



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.

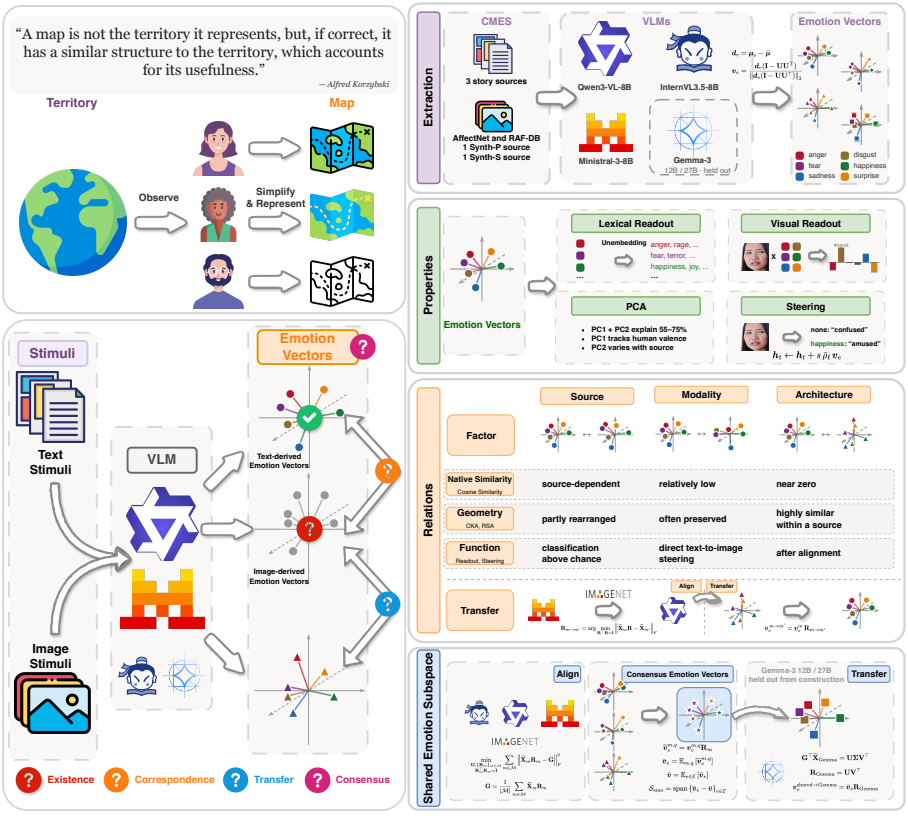



Figure 1: Motivation and study overview. Left: The map–territory analogy and our motivation. Right: Emotion-vector analysis across sources, modalities, and architectures, shared-subspace construction, and held-out Gemma-3 evaluation.

their relational geometry and causal effects generalize across architectures. Further related work is discussed in Appendix A.

As shown on the left side of Figure 1, these gaps lead to four questions. 1) Existence. Do images also produce structured emotion vectors, and how does their geometry depend on stimulus source? 2) Correspondence. Do text- and image-derived vectors correspond, and can text-derived vectors steer image interpretation? 3) Transfer. Do these relations and causal effects extend across architectures? 4) Consensus. Can we construct a shared emotion subspace that preserves this structure and function and generalizes to a held-out architecture? The map–territory analogy summarizes our underlying idea. Different internal representations may take different forms while preserving some common structure.

The right side of Figure 1 summarizes our workflow. To address these questions, we first construct CMES (Cross-Modal Emotion Stimuli), a collection of text and image stimuli organized around the same emotion categories. We use Ekman's six basic emotions, anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992), which are shared by the facial-expression datasets in CMES. Following the story-based protocol of prior work (Sofroniew et al., 2026), CMES includes three independently generated text-story sources. It also includes two real facial-expression datasets (Mollahosseini et al., 2019; Li et al., 2017), synthetic facial portraits, and synthetic emotion-evoking scenes generated with Qwen-Image (Wu et al., 2025). Multiple sources within each modality help distinguish source effects from modality effects. We perform the core analyses with Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), and Ministral-3-8B (Liu et al., 2026). Further details are provided in Appendix B.

Our results answer these questions in sequence. *First*, image inputs produce structured emotion vectors that can be interpreted through lexical and visual readouts. Their low-dimensional structure broadly resembles human valence–arousal organization. PC1 shows relatively stable alignment

with human valence across sources, while PC2's alignment with arousal varies more. Across the three core models, changing stimulus source affects the six-emotion geometry more than changing architecture. *Second*, text- and image-derived emotion vectors show cross-modal correspondence despite low cosine similarity. Before alignment, matched text–image vectors within the same model have cosine similarity of only $0.15$–$0.31$. Yet an alignment fitted on five emotions often generalizes to the held-out sixth in leave-one-emotion-out (LOEO) evaluation. Text-derived vectors also selectively steer the interpretation of face images when injected at image-token positions. Their effects are similar to those of image-derived vectors from the same model. The contrast between low cosine similarity and preserved steering effects motivates us to test whether emotion vectors can also transfer across architectures. *Third*, emotion vectors from different models also correspond after alignment. Matched Ministral-3 and Qwen3-VL emotion vectors have near-zero cosine similarity in their native residual spaces, but LOEO correspondence remains well above permutation controls. A transformation fitted to paired activations on neutral-face or generic ImageNet images recovers substantial correspondence without using the six emotion vectors or their labels. The mapped vectors also transfer causal effects across models. *Finally*, after aligning the three core architectures, we construct a shared emotion subspace from the centered consensus emotion vectors. It preserves the broad valence–arousal organization of the six emotions. The consensus vectors produce selective steering, and their lexical readouts generally show words related to the corresponding emotions. Although Gemma-3-12B and Gemma-3-27B are excluded from consensus construction, the mapped consensus vectors show correspondence with their emotion vectors and causally steer both variants despite their different widths and depths. Together, these results suggest that emotion representations can share relational structure and causal effects across sources, modalities, and architectures. In the map–territory view, different VLMs form different internal maps that preserve part of the same emotion structure. Our main contributions are summarized as follows:

- We introduce CMES, a collection of seven text and image sources covering the same six basic emotions. It supports comparisons of emotion representations and their causal effects across stimulus sources, modalities, and architectures.
- We identify structured image-derived emotion vectors with a low-dimensional organization similar to text-derived vectors. Their geometry varies more across sources than across architectures, while valence alignment is more stable than arousal alignment.
- We show that text- and image-derived emotion vectors within the same model can retain correspondence and similar steering effects despite low cosine similarity.
- We show that mappings learned from paired ImageNet images recover cross-architecture correspondence and enable causal transfer despite near-zero native cosine similarity.
- We construct a shared emotion subspace across three architectures. Its consensus vectors retain the original vectors' functional properties and generalize to held-out Gemma-3.

## 2 CMES AND EMOTION-VECTOR EXTRACTION

CMES covers six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), denoted by $\mathcal{E}$, across seven sources. These include stories generated by Gemini (T1), Qwen3-Max (T2), and Qwen3-VL (T3), real faces from AffectNet and RAF-DB, and synthetic portraits (Synth-P) and synthetic emotion-evoking scenes (Synth-S). Real-face vectors use training images, while functional evaluations use separate validation or test splits.

For each model and source, we use residual activations at approximately two-thirds of the model's depth unless otherwise stated. For text, we average activations over story tokens from position 50 onward, following Sofroniew et al. (2026). For images, we average over all visual tokens. We then average across examples of each emotion $e \in \mathcal{E}$ to obtain its class mean $\boldsymbol{\mu}_e$. We subtract the mean across the six emotions, then remove background directions estimated from neutral stimuli and normalize to obtain

$$\boldsymbol{d}_e = \boldsymbol{\mu}_e - \frac{1}{|\mathcal{E}|}\sum_{e' \in \mathcal{E}} \boldsymbol{\mu}_{e'}, \qquad \boldsymbol{v}_e = \frac{\boldsymbol{d}_e(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)}{\|\boldsymbol{d}_e(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\|_2}, \tag{1}$$

where $\mathbf{U}$ contains the leading neutral PCA (Jolliffe & Cadima, 2016) directions needed to explain at least $50\%$ of the variance. This gives six vectors per model and source. Full setup and protocols are provided in Appendix B.

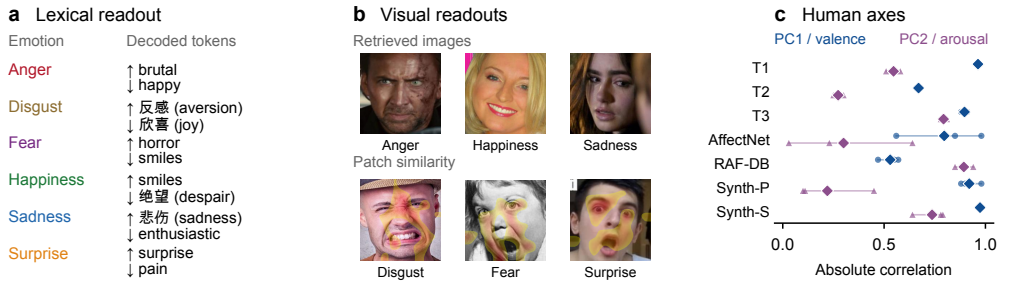



Figure 2: Emotion-vector readouts and human-axis agreement. (a) Selected promoted (↑) and suppressed (↓) tokens for Qwen3-VL AffectNet vectors, with English translations. (b) Selected Qwen3-VL AffectNet retrievals (top) and patch-readout examples with the top 30% of patch similarities highlighted (bottom). (c) Absolute correlations with Russell–Mehrabian (R&M) ratings (Russell & Mehrabian, 1977). Small markers show models, diamonds show means, and lines show ranges.

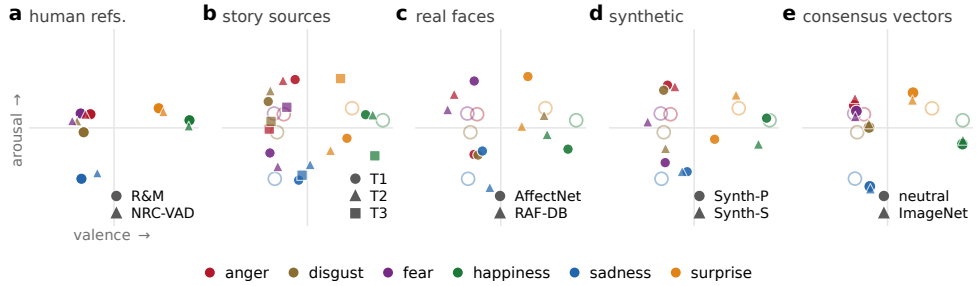



Figure 3: Leading geometry of the six emotion vectors. (a) Human valence–arousal references from R&M and the NRC Valence–Arousal–Dominance (NRC-VAD) lexicon (Mohammad, 2018). (b–d) Qwen3-VL PCA projections from stories, real faces, and synthetic images. (e) Three-model consensus vectors. Faint circles mark R&M positions. Display alignment is described in Appendix B.7.

## 3 Emotion Geometry and Representational Correspondence

### 3.1 Visual Emotion Vectors Form Structured Low-Dimensional Representations

We first explore whether image-derived emotion vectors are interpretable through lexical and visual readouts. We pass each $\boldsymbol{v}_e$ through the model's final normalization and unembedding, then decode the highest- and lowest-scoring tokens. High-scoring tokens generally describe the target emotion, while low-scoring tokens often express opposite emotions, as illustrated in Figure 2a. Figure 2b shows that ranking images by cosine similarity between their pooled activations and each emotion vector retrieves examples of the corresponding class, while patch-level similarity often highlights expressive facial regions within the images. More details are provided in Appendices C.1 and C.2.

We next use PCA to explore the main dimensions of variation among the six emotion vectors. Across the model–source sets, PC1 explains 30–46% of the variance, and PC1+PC2 explain 55–75%. We then correlate the six emotions' PC1 and PC2 coordinates with human valence and arousal ratings, respectively. As shown in Figure 2c, agreement with human valence is generally stronger and more stable, whereas agreement with human arousal varies more across stimulus sources. This variation is not simply a text–image difference, as both modalities include sources with weak and strong arousal agreement. Figure 3 shows how the two-dimensional emotion configurations vary across text and image sources, while the consensus configuration introduced in Section 6 appears closer to the human reference. Full comparisons across vector sets and human ratings are provided in Appendix C.4.

### 3.2 Stimulus Source Changes Relational Geometry More Than Architecture

The PCA visualization shows only the leading two-dimensional plane. We therefore use linear CKA (Kornblith et al., 2019) and RSA (Kriegeskorte et al., 2008) to compare the relationships

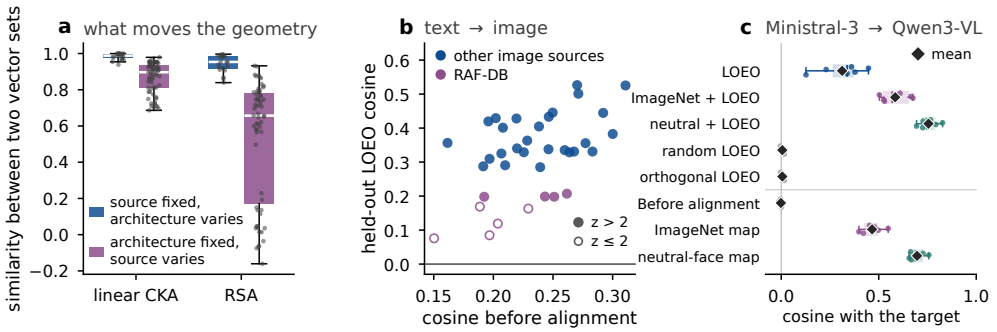



Figure 4: Representational correspondence. (a) CKA and RSA with source or architecture fixed. (b) Text–image cosine before alignment versus LOEO. Filled markers indicate permutation $z > 2$; open markers indicate $z \leq 2$. (c) Ministral-3→Qwen3-VL. Dots show source means before alignment and with LOEO, individual AffectNet emotions after mapping, and source/target replacement means for controls. Diamonds show means; boxes show quartiles and medians.

among all six emotion vectors in the full representation space. To assess how source and architecture affect these relationships, we compare pairs that differ in only one of these factors. As shown in Figure 4a, linear CKA ranges from $0.94$ to $1.00$ with median $0.99$ for the same source across different architectures. For different sources within the same architecture, CKA ranges from $0.69$ to $0.98$ with median $0.90$. The corresponding RSA medians are $0.95$ and $0.66$. Thus, within the three architectures studied here, changing stimulus source affects relational geometry more than changing architecture. RAF-DB accounts for much of the strongest source-specific change. Across the three architectures, its vectors place disgust closer to happiness and fear closer to anger. AffectNet and the other sources more often associate disgust with anger and fear with surprise, consistent with confusion patterns reported in human facial-expression recognition (Mo et al., 2021). The ground-truth labels of retrieved images show similar differences between AffectNet and RAF-DB (Table 4 in Appendix C.2). Together, these results suggest that stimulus source changes the relationships among some emotions, rather than simply weakening emotion information. Detailed geometry and retrieval comparisons are provided in Appendices C.2, C.5, and D.1.

### 3.3 Cross-Modal Correspondence Survives Low Cosine Similarity

We next explore the similarity between text- and image-derived vectors for the same emotion within each VLM. These vectors share residual coordinates, yet their matched-emotion cosine similarity is only $0.15$–$0.31$ across the 36 source/model comparisons. Together with the above geometric results, this suggests that cosine similarity may not fully capture cross-modal correspondence, even within the same model, motivating our cross-modal steering experiments in Section 4. For a stricter test of generalization, we use LOEO alignment to fit an orthogonal transformation (Schönemann, 1966) on five matched emotions and evaluate the sixth. As shown in Figure 4b, held-out cosine ranges from $0.29$ to $0.53$ for AffectNet and the two synthetic image sources. All 27 comparisons exceed our permutation-based threshold of $z = 2$. This suggests that the vectors preserve relationships among emotions across modalities, allowing alignment learned from five emotions to generalize to the sixth. RAF-DB shows weaker LOEO correspondence with both text and other image sources. Text–image LOEO cosine ranges from $0.08$ to $0.21$, with only four of nine comparisons exceeding the permutation threshold, whereas all image–image comparisons exceed it. This weaker correspondence may reflect the different relationships among emotions observed for RAF-DB in the previous subsection. Overall, even when cosine similarity before alignment is low, a mapping learned from five emotions can generalize to the sixth, suggesting that text- and image-derived vectors preserve relationships among emotions. More details are provided in Appendix D.2.

### 3.4 Cross-Architecture Correspondence Survives Near-Orthogonal Native Coordinates

We further explore whether emotion-vector correspondence extends across model architectures. Figure 4c shows the Ministral-3-to-Qwen3-VL comparison as an example. Across the seven stimulus

sources, cosine similarity before alignment stays near zero. In contrast, LOEO cosine ranges from $0.13$ to $0.45$ with mean $0.31$. Matched-norm random and orthogonal LOEO controls instead give near-zero held-out cosine, suggesting that the observed correspondence depends on the emotion relationships rather than on alignment alone. These results suggest that emotion vectors retain similar relationships across models despite their different coordinate bases. A map learned from five emotions can still predict the correspondence for the sixth. This motivates us to explore mappings learned without using the six emotion vectors or their labels. The neutral-face and ImageNet results in Figure 4c preview such mappings, fitted to paired image activations as described in Section 5. Across the six AffectNet emotion pairs, neutral-face mapping gives cosine of $0.66$–$0.76$ with mean $0.70$, while ImageNet mapping gives $0.40$–$0.55$ with mean $0.47$. Together, these results suggest that mappings learned from activations of generic images can recover cross-model emotion correspondence without fitting to the six emotion vectors. More details are provided in Appendix D.3.

# 4 Causal Function Across Modalities

Representational correspondence alone does not show that the vectors affect model behavior. We therefore intervene on activations while each VLM reads held-out face images. At layer $\ell$, we add a scaled unit emotion vector at every image-token position according to

$$\boldsymbol{h}_t \leftarrow \boldsymbol{h}_t + s\,\bar{\rho}_\ell\,\boldsymbol{v}_e, \qquad t \in \mathcal{I}_{\text{img}}, \quad e \in \mathcal{E}, \tag{2}$$

where $\boldsymbol{h}_t$ is the residual activation at token position $t$ in layer $\ell$, $\bar{\rho}_\ell$ is the mean residual norm at that layer, and $s$ controls intervention strength and sign. Unless stated otherwise, we intervene at the same layer $\ell$ used to extract $\boldsymbol{v}_e$, with $s = +0.5$ for positive interventions and $s = -0.5$ for negative interventions. Relative to the unsteered model, we measure the change in log-probability of the injected emotion (*diag*) and the mean change over the five competing emotions (*off*). Positive selective steering should increase the injected-emotion readout and suppress competitors, with the pattern reversing for negative $s$.

## 4.1 Emotion Vectors Selectively Steer Image Interpretation

We first explore whether image-derived emotion vectors can selectively steer image interpretation. As shown in Figure 5a, injection in Qwen3-VL on AffectNet increases the matched emotion's log-probability by $4.96$ and decreases that of competing emotions by $1.43$ on average, with the pattern reversing under negative injection. InternVL3.5 and Ministral-3 show a similar pattern, as shown in Figure 5b. In contrast, matched-norm random and emotion-orthogonal directions produce near-zero changes and do not reproduce the separation between matched and competing emotions. More details are provided in Appendix E.1.

## 4.2 Text-Derived Vectors Functionally Substitute for Image-Derived Vectors

We next replace the image-derived vectors with text-derived vectors from the same model, keeping all other settings unchanged. Despite text–image cosine similarity near $0.2$–$0.3$, the two vector types produce similar steering effects, as shown in Figure 5b. The two vector types also produce similar effects in InternVL3.5 and Ministral-3 and suppress competing emotions in all three models. Detailed results are provided in Appendix E.2. Together with the results in Section 3, these findings suggest that high cosine similarity is not required for either similar relationships among emotions or similar causal effects across modalities.

Additionally, we test whether steering changes explanations beyond the first emotion word while preserving coherent output. As shown in Figure 5c, continuations describe visual cues in ways that match the injected emotion. To test selectivity, we also compare emotion and gender readouts in Qwen3-VL. As shown in Figure 5d, steering changes the emotion prediction on $47\%$ of evaluated RAF-DB images on average, compared with $2\%$ for gender. Steering also shifts the related safe–dangerous prediction toward dangerous for anger and fear and toward safe for happiness, as shown in Figure 5e. To further test whether these vectors contribute to normal emotion recognition, we remove the centered span of the six source-specific vectors. This impairs recognition substantially more than removing a matched-rank random subspace (Table 21 in Appendix E.5). Together, these

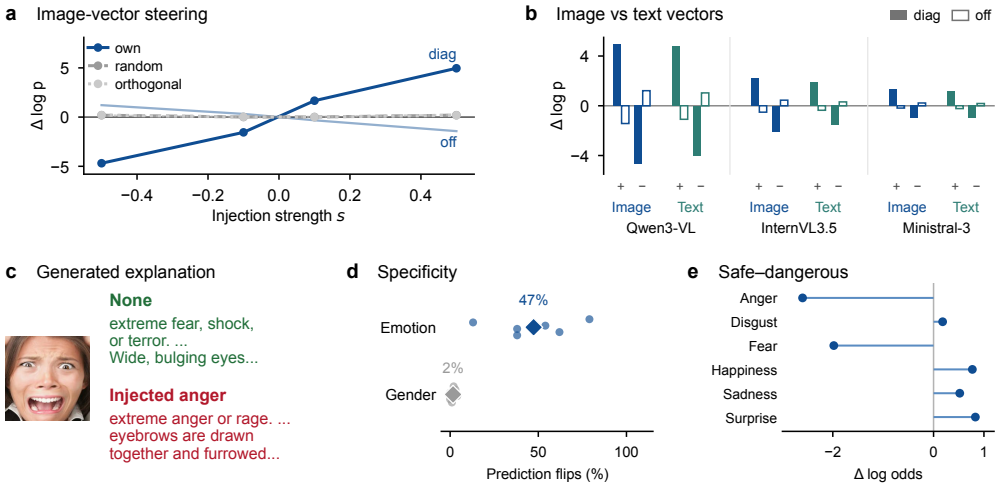



Figure 5: Causal effects within models. (a) Qwen3-VL steering and controls. (b) Image- versus T3 text-derived vectors under positive (+) and negative (−) injection, with filled/open bars showing diag/off. (c) Qwen3-VL continuation excerpts before and after anger injection. (d) Emotion and gender prediction flips across six injections (dots) and their means (diamonds). (e) Changes in safe–dangerous log odds, with positive values favoring safe.

tests suggest that emotion vectors selectively affect how models interpret emotions in images. More details are provided in Appendices E.4, E.5, and E.6.

## 5 Cross-Architecture Transfer

The preceding experiments suggest that emotion vectors have similar relationships and produce similar causal effects across modalities within a model. The cross-architecture LOEO results in Section 3 also provide initial evidence that a mapping fitted directly to five matched pairs of emotion vectors can generalize to the sixth emotion. We therefore explore whether we can learn such a mapping without using emotion information, and whether it can also support causal transfer across architectures.

### 5.1 Estimating the Coordinate Transformation Without the Six Emotion Vectors

We explore whether activations from generic ImageNet images can align different models without using images from emotion datasets. For comparison, we also use activations from neutral faces in an emotion dataset. Let $\widetilde{\mathbf{X}}_m$ and $\widetilde{\mathbf{X}}_{m'}$ contain centered activations from source model $m$ and target model $m'$, with each row pair obtained from the same image. Then, we fit an orthogonal transformation and apply it to emotion vectors from each stimulus source $q$ as follows:

$$\mathbf{R}_{m\to m'} = \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \left\|\widetilde{\mathbf{X}}_m\mathbf{R} - \widetilde{\mathbf{X}}_{m'}\right\|_F, \qquad \boldsymbol{v}_e^{m\to m',q} = \boldsymbol{v}_e^{m,q}\mathbf{R}_{m\to m'}. \tag{3}$$

As shown in Figure 4c, the Ministral-3→Qwen3-VL mapping learned from 3,000 ImageNet validation images raises mean matched-emotion cosine on AffectNet from near zero to 0.47. For comparison, a mapping learned from the same number of neutral faces reaches 0.70. The recovery with ImageNet suggests that generic visual inputs can support cross-model emotion correspondence without using images from emotion datasets. Additional LOEO alignment further improves held-out correspondence for image-derived vectors (Figure 4c) and text-derived vectors (Table 13 in Appendix D.3).

### 5.2 Mapped Emotion Vectors Transfer Causally Across Architectures

We next test whether this recovered correspondence supports causal transfer. As shown in Figure 6, unmapped Ministral-3 image-derived vectors produce essentially no selective effect in Qwen3-

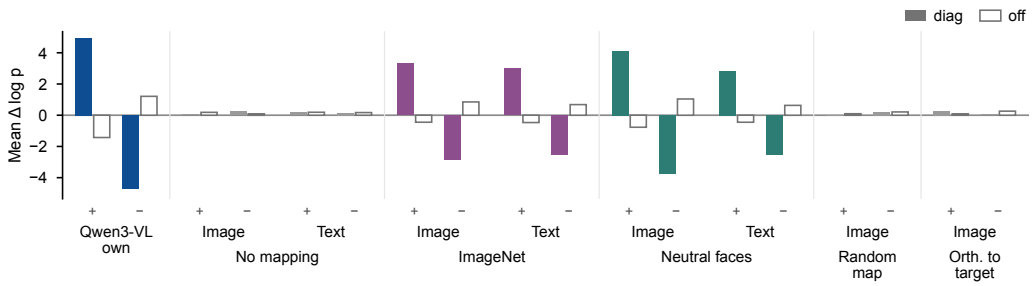



Figure 6: Cross-architecture causal transfer on AffectNet. Ministral-3 image and T2 text vectors are injected into Qwen3-VL, with Qwen3-VL's own image vectors as a reference. Each condition shows positive (+) and negative (−) injections. Filled/open bars show diag/off effects.

VL on AffectNet. Neutral-face and ImageNet mappings restore selective steering, increasing the matched emotion's log-probability by 4.11 and 3.33, respectively, while suppressing competing emotions. We further test transfer across both architectures and modalities by applying these mappings learned from images. For model pairs involving Ministral-3, mapped text vectors also produce selective bidirectional steering despite cosine similarity of only 0.18–0.29. Figure 6 shows this transfer for Ministral-3→Qwen3-VL using T2 vectors. This suggests that mappings learned from image activations can also transfer the causal effects of vectors extracted from text. More details are provided in Appendix E.3.

Together, these results suggest that differences in model coordinates can hide similarities between emotion vectors and limit their direct transfer. A mapping learned from paired image activations can recover both vector correspondence and selective causal effects without fitting to the six emotion vectors or their labels. In addition, these mappings further improve LOEO results, providing more evidence of shared relationships among emotions across models.

## 6 A Shared Emotion Subspace

The above analyses compare pairs of models or sources. We now combine them to construct a shared emotion subspace across the three core architectures and seven stimulus sources, then evaluate the structure and causal effects it preserves. We first extend the pairwise alignment in Equation 3 to the set of three core models $\mathcal{M}$ using generalized Procrustes alignment. We jointly fit orthogonal transformations $\mathbf{R}_m$ and a shared template $\mathbf{G}$, which is the mean of the aligned activations across models. We then align the emotion vectors,

$$\min_{\substack{\mathbf{G},\{\mathbf{R}_m\}_{m\in\mathcal{M}} \\ \mathbf{R}_m^\top\mathbf{R}_m=\mathbf{I},\ \forall m\in\mathcal{M}}} \sum_{m\in\mathcal{M}} \left\|\widetilde{\mathbf{X}}_m\mathbf{R}_m - \mathbf{G}\right\|_F^2, \quad \mathbf{G} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \widetilde{\mathbf{X}}_m\mathbf{R}_m, \quad \widetilde{\boldsymbol{v}}_e^{m,q} = \boldsymbol{v}_e^{m,q}\mathbf{R}_m. \tag{4}$$

This places all three models in a common frame without choosing one as the reference. We average the aligned vectors over models and sources to obtain the consensus vector $\bar{\boldsymbol{v}}_e$ for each emotion $e \in \mathcal{E}$, then average them across emotions,

$$\bar{\boldsymbol{v}}_e = \mathbb{E}_{m,q}\left[\widetilde{\boldsymbol{v}}_e^{m,q}\right], \qquad \bar{\boldsymbol{v}} = \mathbb{E}_{e\in\mathcal{E}}\left[\bar{\boldsymbol{v}}_e\right]. \tag{5}$$

We then define the shared emotion subspace as

$$\mathcal{S}_{\text{emo}} = \text{span}\left\{\bar{\boldsymbol{v}}_e - \bar{\boldsymbol{v}}\right\}_{e\in\mathcal{E}}. \tag{6}$$

Centering removes the mean, so $\mathcal{S}_{\text{emo}}$ spans the shared vectors that distinguish the six emotions.

We first use variance decomposition to explore how much variation is shared across models and sources. After alignment, shared differences among emotions remain, with little model-wide variation but substantial source variation and interactions. We map $\mathcal{S}_{\text{emo}}$ back into each model's coordinates and compute cosine similarity between each original vector and its orthogonal projection onto this subspace. Under neutral-face alignment, the mean cosine over all seven sources and six emotions ranges from 0.59 to 0.64 across the three models. ImageNet alignment gives mean cosine of 0.57–0.63, along with a similar emotion share in the variance decomposition. More details are provided in Appendix F.1.

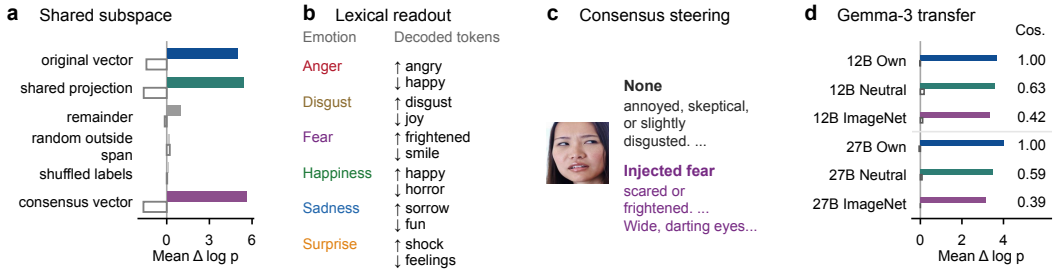



Figure 7: Shared-subspace effects and consensus transfer. (a) Qwen3-VL interventions. (b) Selected promoted (↑) and suppressed (↓) tokens for Ministral-3 under ImageNet alignment. (c) Qwen3-VL continuations before and after consensus fear injection. (d) Gemma's own vectors versus neutral-face/ImageNet-mapped consensus vectors at $s = +0.1$. Filled/open bars in (a, d) show diag/off. Cos. denotes mean cosine similarity to the corresponding Gemma model's own AffectNet vectors.

We next test whether the part of each original vector inside the shared subspace retains its causal effects. We split each Qwen3-VL emotion vector into this shared projection and the orthogonal remainder outside $\mathcal{S}_{\text{emo}}$, then inject each part separately (Equation 26 in Appendix F.2). As shown in Figure 7a, the shared projection increases the matched emotion's log-probability slightly more than the original vector. The remainder produces only a very small effect. Random directions outside the shared subspace and remainders with shuffled emotion labels show little selective steering. In addition, we map the consensus vectors $\bar{\boldsymbol{v}}_e$ back into Qwen3-VL's coordinates and find that they produce selective steering comparable to its original vectors. More details are provided in Table 26 in Appendix F.2.

We also explore whether the consensus vectors retain lexical meaning and steer generated continuations. When mapped back to each model's native coordinates, consensus vectors show words related to the target emotion in their lexical readouts, as shown in Figure 7b. Figure 7c shows that Qwen3-VL continuations also follow the injected emotion and describe supporting visual cues. Together, these examples suggest that consensus vectors can produce clearer emotion-related readouts and steer continuations more effectively than the original emotion vectors. Figure 3e also illustrates their agreement with the human valence–arousal reference under both alignment settings. To further test whether the consensus extends beyond the models used to construct it, we evaluate Gemma-3-12B and include Gemma-3-27B as a model-size ablation. Neither contributes to consensus construction, yet the mapped consensus vectors show correspondence with each variant's own vectors and retain causal effects in both, as shown in Figure 7d. These results support a shared emotion subspace whose consensus vectors retain lexical meaning and transfer causal effects to a held-out architecture at both model sizes. More details are provided in Appendices F.3 and F.4.

## 7 Conclusions

Our results suggest that emotion concepts in VLMs generalize through shared relationships among emotions and transferable causal effects, even when individual vector directions differ across sources, modalities, and architectures. Within a model, text- and image-derived vectors show held-out correspondence despite low cosine similarity, and text-derived vectors directly steer image interpretation. Across architectures, held-out correspondence can persist even when cosine similarity is near zero. Mappings learned from paired image activations then recover vector correspondence and causal transfer without using emotion information. Cosine similarity before alignment therefore gives an incomplete picture of this generalization. We further construct a shared emotion subspace across models and sources. This subspace preserves agreement with human valence and arousal, while projecting Qwen3-VL's original vectors onto it retains most of their selective steering effect. Qualitative examples also suggest that the consensus vectors produce clearer emotion-related lexical readouts and steer continuations more effectively than the original vectors. Moreover, they show correspondence and causal effects in Gemma-3-12B and Gemma-3-27B, although neither model contributes to their construction. We therefore use the map–territory analogy to summarize our findings: VLMs form different internal maps of emotion, yet these maps preserve some common structure and transferable causal effects.

## REFERENCES


Ayushi Agarwal. Cross-architecture steering transfer in language models: A systematic empirical study, 2026. URL https://arxiv.org/abs/2608.05164.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48531–48563. Curran Associates, Inc., 2025. doi: 10.52202/085713-1621. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4569a868e7aa891248832ec08445d071-Paper-Conference.pdf.

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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.185. URL https://aclanthology.org/2025.acl-long.185/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.

Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, apr 2016. ISSN 2050-084X. doi: 10.7554/eLife.10989. URL https://doi.org/10.7554/eLife.10989.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 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.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593, 2017. doi: 10.1109/CVPR.2017.277.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lian25a.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Fan Mo, Jingjin Gu, Ke Zhao, and Xiaolan Fu. Confusion effects of facial expression recognition in patients with major depressive disorder and healthy controls. *Frontiers in Psychology*, 12:703888, 2021. doi: 10.3389/fpsyg.2021.703888. URL https://doi.org/10.3389/fpsyg.2021.703888.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1017. URL https://aclanthology.org/P18-1017/.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild . *IEEE Transactions on Affective Computing*, 10(01):18–31, January 2019. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2740923. URL https://doi.ieeecomputersociety.org/10.1109/TAFFC.2017.2740923.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002. ISSN 1097-0193. doi: 10.1002/hbm.1058.

Maya Okawa, Bo Zhao, Eric J. Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models, 2026. URL https://arxiv.org/abs/2507.10599.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. ISSN 0954-5794. doi: 10.1017/S0954579405050340.

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=72TN9UAtNI.

James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977. ISSN 0092-6566. doi: https://doi.org/10.1016/0092-6566(77)90037-X. URL https://www.sciencedirect.com/science/article/pii/009265667790037X.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. ISSN 1860-0980. doi: 10.1007/BF02289451.

Marie L. Smith, Garrison W. Cottrell, Frédéric Gosselin, and Philippe G. Schyns. Transmitting and decoding facial expressions. *Psychological Science*, 16(3):184–189, 2005. ISSN 0956-7976. doi: 10.1111/j.0956-7976.2005.00801.x.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Sinie van der Ben, Raphaël Baur, Yannick Metz, and Mennatallah El-Assady. Where do models find happiness? emotion vectors in open-source LLMs, 2026. URL https://arxiv.org/abs/2606.26987.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=a8N0nRG3jA.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. EMO-LLaMA: Enhancing facial emotion understanding with instruction tuning. *International Journal of Computer Vision*, 134(7):350, 2026. doi: 10.1007/s11263-026-02918-9. URL https://doi.org/10.1007/s11263-026-02918-9.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wnKgWnx8mH.

Li Zheng, Xin Zhang, Shuyi He, Fei Li, Chong Teng, Jiang-Ming Yang, Donghong Ji, and Zhuang Li. Are emotion and rhetoric neurons in LLM? neuron recognition and adaptive masking for emotion-rhetoric prediction steering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4206–4216, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.193. URL https://aclanthology.org/2026.acl-long.193/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2025. URL https://arxiv.org/abs/2310.01405.

# APPENDIX

The appendix is organized as follows.

- **Appendix A**. Related work. We review emotion representations, activation steering, and cross-model transfer, and explain how our study differs from prior work.
- **Appendix B**. Experimental setup and reproducibility. This section documents the models, stimulus sources and splits, generation prompts, emotion-vector extraction, steering and alignment protocols, similarity measures, and research materials planned for release.
- **Appendix C**. Existence and characterization across modalities, sources, and models. We present lexical and visual evidence for emotion vectors and explore their structure across depth, agreement with human affective ratings, relational geometry, and sensitivity to stimulus source and filtering.
- **Appendix D**. Representational similarity. We compare matched emotion vectors within modalities, across modalities, and across architectures using cosine similarity before alignment and held-out LOEO alignment with permutation controls. We also test mappings learned from paired image activations.
- **Appendix E**. Causal steering. We test steering within models and transfer across modalities and architectures. Further experiments explore specificity, continuation coherence, injection position, removal of the emotion-vector span, effects beyond the emotion readout, and depth dependence.
- **Appendix F**. Shared emotion subspace. We construct and evaluate the shared subspace through variance decomposition, geometric reconstruction, causal interventions, lexical readouts, and generated continuations. We also test held-out generalization to Gemma-3-12B and include Gemma-3-27B as a model-size ablation.
- **Appendix G**. Limitations, future work, and broader impact. We discuss limitations in model and emotion coverage, source dependence, and the need to evaluate practical applications. We also outline future research directions and potential risks of emotion inference and steering.
- **Appendix H**. Ethics statement. This statement addresses data access and privacy, restrictions on redistributing real face images, limits of emotion inference, and potential misuse of emotion steering.
- **Appendix I**. Reproducibility statement. This statement identifies where experimental protocols, results, and controls are documented and summarizes the planned release of research materials.
- **Appendix J**. LLM usage statement. This statement describes the use of generative models for synthetic stories and images and of LLMs for language polishing.

## A RELATED WORK

**Emotion representations in language models.** Recent work suggests that LLMs contain structured emotion representations. Reichman et al. (2026) report a low-dimensional emotional manifold whose main dimensions resemble human valence and arousal (Posner et al., 2005). More recently, van der Ben et al. (2026) recover similar valence-related geometry in two open-weight LLMs, while finding that arousal alignment is more sensitive to the story corpus used for extraction. Other studies identify emotion directions that generalize across contexts and their associated circuits (Wang et al., 2026), as well as hierarchical organization similar to human emotion taxonomies (Okawa et al., 2026). However, these studies mainly characterize emotion structure from text, leaving its consistency across text and image sources unclear. We extend prior work to image-derived emotion vectors and compare the geometry of the six emotions across stimulus sources, modalities, and architectures.

**Activation steering and causal intervention.** Activation steering changes model behavior by intervening along internal directions (Zou et al., 2025). For emotion, prior work shows that interventions along emotion directions can alter emotional judgments, downstream decisions, and generated

behavior (Sofroniew et al., 2026; Wang et al., 2026; Zheng et al., 2026; Reichman et al., 2026). These studies mainly consider interventions within individual language models using text inputs, providing limited evidence for the transfer of emotion steering across modalities and architectures. More recently, Zhang et al. (2026) trace image-to-text information flow and identify emotion-related attention heads and neurons within individual VLMs. In contrast, we study whether independently extracted text- and image-derived emotion vectors preserve relational geometry and causal effects across sources, modalities, and architectures.

**Cross-model representational and steering transfer.** Prior work has shown that representations learned by different neural networks can often be related despite differences in their native coordinate systems (Bansal et al., 2021; Moschella et al., 2023; Huh et al., 2024). More recent work directly aligns concept representations across text LLMs and transfers probes or steering vectors between models using linear or affine transformations (Huang et al., 2025; Chen et al., 2025; Agarwal, 2026). However, successful transfer of individual probes or steering vectors does not by itself establish that a group of related concepts retains its internal geometry and causal function across modalities and architectures. We study six related emotion concepts as a group and compare their relational geometry across stimulus sources, modalities, and architectures. We also evaluate the transfer of their causal effects across models after alignment, linking the organization of the emotion group to the function of its individual vectors.

# B Experimental Setup and Reproducibility

This appendix specifies the models, stimulus sources, activation-extraction procedure, intervention protocol, cross-architecture alignment, and similarity measures used throughout the paper. Unless depth is varied explicitly, analyses use layers at approximately two-thirds of each model's depth. Representational analyses use layer 24 in all three core models. For causal experiments, emotion vectors are extracted and injected at layer 24 in Qwen3-VL and InternVL3.5 and layer 23 in Ministral-3. We study the six Ekman emotions—anger, disgust, fear, happiness, sadness, and surprise. Neutral stimuli are used to estimate nuisance directions and are not treated as a seventh emotion vector.

## B.1 Models and Inference

We conduct the core analyses on three instruction-tuned VLMs: Qwen3-VL-8B-Instruct[1] (36 layers, $D = 4096$), InternVL3.5-8B-Instruct[2] (36 layers, $D = 4096$), and Ministral-3-8B-Instruct-2512[3] (34 layers, $D = 4096$). For held-out architecture generalization (Appendix F.4), we additionally evaluate Gemma-3-12B-IT[4] (48 layers, $D = 3840$) and Gemma-3-27B-IT[5] (62 layers, $D = 5376$) as a within-family model-size ablation. The Gemma-3 analyses and interventions use layers 32 and 41 for 12B and 27B, respectively. Both variants differ in residual width and depth from the core models and are excluded from the three-model consensus construction. All experiments use bfloat16 inference on NVIDIA GH200 GPUs with each model's native multimodal processor.

## B.2 Stimulus Sources

CMES contains three text-story sources and four image sources. The text sources are a public story corpus generated by Gemini 3.1 Pro Preview (T1; `https://huggingface.co/datasets/ryancodrai/emotion-probes`) and two story sets generated for this work with Qwen3-Max (T2) and Qwen3-VL-8B (T3). T3 serves as a self-generation control for Qwen3-VL and as a third matched story source for the other architectures. The image sources comprise real facial-expression images from AffectNet and RAF-DB, synthetic facial portraits generated with Qwen-Image (Synth-P), and synthetic emotion-evoking scenes generated with Qwen-Image (Synth-S).

[1] `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`
[2] `https://huggingface.co/OpenGVLab/InternVL3_5-8B-Instruct`
[3] `https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512`
[4] `https://huggingface.co/google/gemma-3-12b-it`
[5] `https://huggingface.co/google/gemma-3-27b-it`

Table 1: Stimulus sources and sample counts. Unfiltered text counts follow the generation protocol. Every set contains six target emotion classes and a neutral class. Per-class counts cover all seven classes; the filtered text block lists counts per target emotion and for neutral separately. QF means Qwen-filtered, retaining stimuli whose labels are correctly predicted by Qwen3-VL.

| Label | Source | Stimulus type | Per class | Total |
|---|---|---|---|---|
| *Text sources* | | | | |
| T1 | Gemini 3.1 Pro Preview | Generated text | 1,200 | 8,400 |
| T2 | Qwen3-Max stories | Generated text | 1,200 | 8,400 |
| T3 | Qwen3-VL stories | Generated text | 1,200 | 8,400 |
| *Image sources* | | | | |
| AffectNet | | Real faces | 3,000 | 21,000 |
| RAF-DB | | Real faces | 281–500 | 3,281 |
| Synth-P | synthetic portraits | Generated images | 3,000 | 21,000 |
| Synth-S | synthetic scenes | Generated images | 3,000 | 21,000 |
| *Success-filtered image controls* | | | | |
| AffectNet (filtered) | | Real faces | 377 | 2,639 |
| RAF-DB (filtered) | | Real faces | 106 | 742 |

| *Success-filtered text controls* | | | | |
|---|---|---|---|---|
| Label | Source | Per emotion | Neutral | Total |
| T1-QF | T1, filtered | 622 | 395 | 4,127 |
| T2-QF | T2, filtered | 188 | 343 | 1,471 |
| T3-QF | T3, filtered | 82 | 506 | 998 |

Each source contains the six Ekman emotions together with a neutral class used for nuisance-component estimation. AffectNet emotion vectors are estimated exclusively from the training split and evaluated on the validation split; RAF-DB vectors are likewise estimated from training and evaluated on test. No evaluation image is used to construct an emotion vector. Success-filtered variants are used only in Appendix C.6.

### B.3 Stimulus Generation and Prompts

Figure 8 shows the prompt templates for generating stories, synthetic portraits, and synthetic scenes.

**Stories.** T2 and T3 follow the story-generation protocol of Sofroniew et al. (2026). The full generation protocol crosses a 171-word emotion vocabulary with 100 topics and requests 12 independent stories per emotion–topic pair. For the six target emotion classes used here, we request 1,200 stories per class and 7,200 in total. Neutral stories are generated separately with the same generator, using a prompt that asks for factual descriptions without emotional content. The conditioning instruction is never supplied to the probed model; activation extraction receives only the generated story text. Among the inspected stories from the six emotion-conditioned classes, the conditioning word or its corresponding noun/adjective form appears in only $0.14\%$ of T2 stories and $0.85\%$ of T3 stories, remaining below $1.5\%$ in every class. We retain these rare occurrences rather than filtering examples based on surface text.

**Synthetic images.** Synthetic images are generated with `Qwen/Qwen-Image-2512` at $1024 \times 1024$ resolution using 50 diffusion steps and one recorded random seed per image. Portrait prompts cross emotion with demographic, style, composition, and lighting. Scene prompts instead use environment-centric topics. No generated image is manually selected, filtered, or regenerated after generation.

### B.4 Activation Extraction and Emotion-Vector Construction

**Text pooling.** For each story, we forward only the generated story text without a chat template, hook the residual stream at every transformer layer, and average activations from token position 50 to the final story token. Stories shorter than 60 tokens are discarded.

**Stories (T2 and T3)**

```
Write {n_stories} different stories based on the following premise.
Topic: {topic}
The story should follow a character who is feeling {emotion}.
Format the stories like so: [...] The paragraphs should each be a
fresh start, with no continuity. Try to make them diverse and not
use the same turns of phrase.
```

**Synthetic portraits (Synth-P)**

```
A {style} {composition} of a {demographic} feeling {emotion}.
{lighting}. Photorealistic, expressive face, no text, no watermark.
```

**Synthetic scenes (Synth-S)**

```
A {style} photograph of {topic}. Atmosphere: {emotion}. {lighting}.
Photorealistic, scene-focused composition, no text, no watermark.
```

Figure 8: Prompt templates for CMES stimulus generation. Braces mark placeholders. The story-format example is omitted.

**Image pooling.** Images are processed with each model's native multimodal processor. We identify the model-specific visual-token span in the processed sequence and average residual activations over that span. The same model-specific span definition is used for subsequent image-token steering.

**Emotion vectors.** For model $m$, source $q$, layer $\ell$, and emotion $e \in \mathcal{E}$, let $\boldsymbol{\mu}_e$ denote the mean pooled activation of that class and

$$\bar{\boldsymbol{\mu}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \boldsymbol{\mu}_e. \tag{7}$$

We form the centered class contrast

$$\boldsymbol{d}_e = \boldsymbol{\mu}_e - \bar{\boldsymbol{\mu}}. \tag{8}$$

Neutral stimuli are used separately to estimate high-variance nuisance directions. Let $\widetilde{\mathbf{X}}_0$ collect centered neutral activations and let $\mathbf{U}$ contain the leading principal components whose cumulative explained variance first reaches $50\%$. The final emotion vector is

$$\boldsymbol{v}_e = \frac{\boldsymbol{d}_e(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)}{\|\boldsymbol{d}_e(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\|_2}. \tag{9}$$

### B.5 Steering Protocol

At intervention layer $\ell$, we add a scaled emotion vector to each residual-stream position in the image-token span:

$$\boldsymbol{h}_t \leftarrow \boldsymbol{h}_t + s\,\bar{\rho}_\ell\,\boldsymbol{v}_e, \qquad t \in \mathcal{I}_{\text{img}}, \tag{10}$$

where $\bar{\rho}_\ell$ is the mean residual norm at that layer estimated from $40$ held-out images. Unless stated otherwise, we use unit-length vectors at one layer with $s = +0.5$, or $s = -0.5$ for negative interventions. Appendix F.2 additionally tests the shared projection and remainder at their original lengths, so the two injected components sum to the original vector before the model response is measured.

The model answers *How does this person feel?* after the prime *They feel*. We score the label words *angry, disgusted, afraid, happy, sad, surprised*, and *neutral* from the next-token distribution. Only the six emotion directions are injected. For an injected emotion $e$, *diag* is the mean change in $\log p(e)$ relative to the unsteered model, while *off* is the mean change over the other five emotion labels. Neutral remains a candidate readout and an input class, but is excluded from the off-diagonal average and is not a seventh emotion vector.

Within-model steering uses held-out images selected as correctly classified in a separate screening run, capped at 50 per class including neutral. This yields $274/283$ RAF-DB/AffectNet images for Qwen3-VL, $249/243$ for InternVL3.5, and $253/227$ for Ministral-3. The Gemma-3 AffectNet subsets contain 283 images for 12B and 332 for 27B, also including neutral inputs. Every intervention run recomputes the unsteered distribution on its evaluation images. Accuracy under additional readout conditions is measured with the relevant candidate labels. Where uncertainty is reported, confidence intervals use 2,000 paired percentile-bootstrap resamples over images, keeping all six injections for each sampled image together.

### B.6 Cross-Architecture Alignment

For source model $m$ and target model $m'$ among the three equal-width core architectures, we estimate an orthogonal transformation from paired activations without using the six emotion vectors or their labels. Given centered paired activation matrices $\widetilde{\mathbf{X}}_m$ and $\widetilde{\mathbf{X}}_{m'}$, we fit

$$\mathbf{R}_{m\to m'} = \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \left\| \widetilde{\mathbf{X}}_m\mathbf{R} - \widetilde{\mathbf{X}}_{m'} \right\|_F. \tag{11}$$

An emotion vector from stimulus source $q$ is transferred by

$$\boldsymbol{v}_e^{m\to m',q} = \boldsymbol{v}_e^{m,q}\mathbf{R}_{m\to m'}. \tag{12}$$

The default transformation is estimated from 3,000 AffectNet neutral training faces processed by both models. As an out-of-domain control, we fit the same transformation from 3,000 generic ImageNet validation images. For the three-model analysis, generalized Procrustes alignment places all architectures in a symmetric consensus frame before the shared emotion subspace is constructed. For the unequal-width Gemma-3 transfers, we compute the thin SVD $\widetilde{\mathbf{X}}_m^\top\widetilde{\mathbf{X}}_{m'} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and use the rectangular map $\mathbf{U}\mathbf{V}^\top$. This map need not preserve every source direction's norm, so transferred vectors are normalized in the target space before standard steering. Gemma-3 does not contribute emotion vectors to consensus construction; its paired activations are used only to estimate a map from the fitted consensus frame to Gemma-3's coordinates.

### B.7 Similarity Measures

Given two six-emotion sets $\{\boldsymbol{v}_e\}_{e\in\mathcal{E}}$ and $\{\boldsymbol{w}_e\}_{e\in\mathcal{E}}$, we distinguish axis alignment, relational geometry, and one-to-one vector correspondence.

**Axis alignment.** For each set, the six vectors are stacked as rows into $\mathbf{V} \in \mathbb{R}^{|\mathcal{E}|\times D}$, centered across emotions, and analyzed with PCA. PC1 and PC2 are oriented using the Russell–Mehrabian reference: PC1 is sign-aligned to pleasure and PC2 to arousal. Corresponding axes are compared using Pearson correlation over the six emotions.

For display in Figure 3, each two-dimensional configuration is centered, normalized to unit Frobenius norm, and Procrustes-aligned to R&M. This display alignment allows rotation and reflection, while the axis correlations above use only sign alignment. Faint circles mark the normalized R&M reference positions.

**Relational geometry.** Linear CKA compares centered inner-product geometry:

$$\mathrm{CKA}(\mathbf{V},\mathbf{W}) = \frac{\left\|\mathbf{W}^\top\mathbf{V}\right\|_F^2}{\left\|\mathbf{V}^\top\mathbf{V}\right\|_F\left\|\mathbf{W}^\top\mathbf{W}\right\|_F}. \tag{13}$$

RSA instead forms each set's $6\times 6$ cosine-distance matrix and computes the Spearman correlation between the 15 unique pairwise distances. CKA and RSA therefore compare relational structure rather than native vector coordinates.

**Vector correspondence.** Mean cosine similarity measures matched-emotion correspondence in the original coordinates as follows.

$$\mathrm{CosSim}(\mathbf{V},\mathbf{W}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\cos(\boldsymbol{v}_e,\boldsymbol{w}_e). \tag{14}$$

Full-fit cosine fits an orthogonal Procrustes transformation using all six matched emotions and evaluates those same pairs, so it is treated only as an optimistic description of alignability. LOEO cosine instead fits on five matched emotions and evaluates the held-out sixth, averaging over all six folds.

For each fit, the training rows are centered separately in the two spaces. Because five emotion vectors do not determine a full residual-space rotation, we select the orthogonal least-squares solution closest to the identity among solutions with the same fitting error. This specifies the otherwise undetermined action outside the fitted directions and preserves vector norms in the common fitting space. For unequal widths, we first pad the shorter space with zeros, apply the same square orthogonal fit, and retain the target coordinates for cosine evaluation. The padding is a computational convention and does not define native cosine across unequal widths. These emotion-based fits are separate from the activation-based transformations in Appendix B.6.

The held-out emotion pair is excluded from alignment fitting, while its vectors follow the common six-class construction in Appendix B.4. LOEO therefore tests generalization of the relation among the six emotion vectors, rather than extraction of a new emotion concept without any examples from that class.

**Permutation null.** To test whether LOEO correspondence is specific to emotion identity, we repeat the procedure after permuting the labels of one set (Nichols & Holmes, 2002) and standardize the observed score as

$$z = \frac{\mathrm{LOEO}_{\mathrm{obs}} - \mu_{\mathrm{perm}}}{\sigma_{\mathrm{perm}}}. \tag{15}$$

The null uses $500$ non-identity label shuffles drawn with a fixed random seed. Each shuffle reruns all six LOEO folds, including fitting each transformation on the other five emotion pairs. Thus, the null evaluates the same held-out statistic as the observed score. We use $z > 2$ as evidence that correspondence exceeds that expected from the unlabeled six-point geometry and report cosine alongside this standardized score to show its magnitude.

### B.8 Reproducibility Artifacts

We will release the analysis code, complete generation prompts, stimulus sample lists, synthetic-generation seeds, experiment configurations, and extracted emotion vectors. Exact model revisions, software environments, and job scripts are provided with the artifact. RAF-DB and AffectNet images themselves are not redistributed; instead, we release the identifiers required to reconstruct the experimental subsets under their original licences.

## C Existence and Characterization Across Modalities, Sources, and Models

This appendix explores what the emotion vectors represent and how the six emotions are related across stimulus sources and models. We first explore the words, images, and image regions associated with individual vectors in Appendices C.1 and C.2. We then compare the relationships among the six vectors across layers in Appendix C.3, with human valence and arousal in Appendix C.4, and across sources and models in Appendix C.5. Finally, Appendix C.6 explores how stimulus source and selection affect the leading axes. These analyses provide the basis for the correspondence, causal steering, and shared-subspace experiments in Appendices D–F.

### C.1 Logit-Lens of Emotion Vectors

We first explore whether the emotion vectors promote words related to their target emotions. Following Sofroniew et al. (2026), we pass each vector through the model's final normalization and unembedding and inspect its highest- and lowest-scoring tokens. Figure 9 compares text-story and AffectNet image vectors at layers 18 and 24. These are at $0.50$ and $0.67$ of the depth of Qwen3-VL and InternVL3.5, and at approximately $0.53$ and $0.71$ of the depth of Ministral-3. At layer 18, text vectors already tend to promote emotion-related words, although the pattern is less clear for Ministral-3. Image vectors at the same layer also promote words describing visual content. For

Qwen3-VL

| | Text, layer 18 | Text, layer 24 | Image, layer 18 | Image, layer 24 |
|---|---|---|---|---|
| Anger | ↑ 毁灭, angrily, merciless, 狠<br>↓ 清凉, sweetness, sweet, 清爽 | ↑ angrily, 侮辱, 指责, 怒<br>↓ 午后, pleasantly, 安心, 香味 | ↑ 西装, 父子, 的妻子, 威名<br>↓ 宝宝, gigg, 甜美, 活泼 | ↑ 强悍, 霸气, 强硬, brutal<br>↓ 喜悦, 笑声, happy, joy |
| Disgust | ↑ 肮, disgusting, 毒, 污<br>↓ tapping, rhyth, woven, heartbeat | ↑ disgusting, 毒, 呕吐, disgust<br>↓ melody, 旋律, heartbeat, relaxed | ↑ 搞笑, 嘴, 詈, /problem<br>↓ 眷, 绘本, beautifully, 邂 | ↑ 嘴, 眉头, 不舒服, 反感<br>↓ 杳, 眷, 绘本, 照亮 |
| Fear | ↑ 畏惧, 恐惧, 害怕, 不敢<br>↓ 欢快, 幽默, 趣, 废话 | ↑ 害怕, fearful, 恐惧, 畏惧<br>↓ 幽默, 有趣的, 调侃, 有趣 | ↑ Horror, 惊, screams, screaming<br>↓ 雅, 典雅, 退休, retirement | ↑ 惊, 震惊, horror, 惊讶<br>↓ 微笑, smiles, prosper, 低调 |
| Happiness | ↑ playful, delightful, mischief, joyful<br>↓ 不堪, 畏惧, 惨, Worse | ↑ joyful, playful, mischief, delighted<br>↓ 惨, 残酷, painfully, Worse | ↑ 愉快, 开朗, 明媚, warmly<br>↓ 不堪, 动作, rypt, 重重 | ↑ smiles, smiling, 微笑, smile<br>↓ 绝望, 哭泣, 不堪, 恐惧 |
| Sadness | ↑ 静静, 寂, 孤独, 眺<br>↓ ！！, !!!, !!, !( | ↑ 寂寞, 孤单, 孤独, melanch<br>↓ freaking, 绝对是, Seriously, 当场 | ↑ 哭泣, 流泪, 泪水, 悲伤<br>↓ enthusiastically, 主持, 蹿, hype | ↑ 哭泣, 泪水, 悲伤, 泪<br>↓ enthusiastically, enthusiastic, 夸张, enthusiasm |
| Surprise | ↑ 突然, suddenly, 突如其, Suddenly<br>↓ 詈, pretended, 大家分享, 精神病 | ↑ Suddenly, suddenly, uddenly, 突如其来<br>↓ pretended, 报复, 出去, 后来 | ↑ 呱, 宝宝, 绘本, anomal<br>↓ vengeance, 疤, 钢筋, 意志 | ↑ 好奇, 惊讶, surprise, 惊奇<br>↓ 泪水, 苦难, pain, 折磨 |

InternVL3.5

| | Text, layer 18 | Text, layer 24 | Image, layer 18 | Image, layer 24 |
|---|---|---|---|---|
| Anger | ↑ 破坏, 痛, 崩溃, 报废<br>↓ 清澈, serene, noticing, 温暖 | ↑ 破坏, 怒, 精确, 毁<br>↓ somehow, warmth, peek, nestled | ↑ 西装, 父子, 退休, imposing<br>↓ surprise, 宝宝, bubble, 乐园 | ↑ 对抗, 强悍, 统治, cret<br>↓ 快乐, funny, 惊喜, joyful |
| Disgust | ↑ afterwards, 垃圾, 毒, 彻底<br>↓ woven, heartbeat, rhyth, light | ↑ 毒, 呕吐, toxic, 彻底<br>↓ relaxed, melody, paused, heartbeat | ↑ 劲, 试图, 嘴, 战术<br>↓ 清澈, beautifully, document, fresh | ↑ 皱, 眉头, 尝试, 扭曲<br>↓ 清澈, clean, 杳, surrounded |
| Fear | ↑ constantly, imagined, 恐惧, lest<br>↓ 各自, 协同, 还原, 废话 | ↑ 害怕, 恐惧, fear, 不安<br>↓ 有趣的, 大致, 幽默, 休闲 | ↑ action, reaction, situations, abnormal<br>↓ confident, 典雅, 自信, 定居 | ↑ 惊, 瞬间, 反应, suddenly<br>↓ confident, 微笑, smiling, modest |
| Happiness | ↑ playful, joy, 俏, harmony<br>↓ 病理, Worse, ako, sterile | ↑ playful, joyful, joy, mischief<br>↓ Worse, negatively, 缩水, dissoci | ↑ smiling, smiles, smile, neatly<br>↓ 不堪, 加剧, 防御, 抗议 | ↑ smiling, smile, 微笑, smiles<br>↓ 加剧, 迫使, 警告, 危险 |
| Sadness | ↑ unchanged, 孤独, 依旧, still<br>↓ 绝对是, 立即, 最关键, ！ | ↑ still, 孤独, lonely, unchanged<br>↓ 绝对是, ！, 警告, 当场 | ↑ 不堪, 哭泣, unacceptable, 无助<br>↓ ener, enthusiasts, enthusiastic, 组合 | ↑ 哭泣, 悲伤, 流泪, 泪水<br>↓ enthusiastic, enthusiastically, marketing, hype |
| Surprise | ↑ suddenly, 突然, Suddenly, now<br>↓ 熵, pretended, 各自, 各自的 | ↑ suddenly, Suddenly, 难道, 忽然<br>↓ osti, 棍, 证明, 滚 | ↑ surprise, 问我, cute, 宝宝<br>↓ 皱纹, 意志, 不堪, 妥协 | ↑ 好奇, surprise, 惊喜, curious<br>↓ 皱纹, ful, tear, 摧 |

Ministral-3

| | Text, layer 18 | Text, layer 24 | Image, layer 18 | Image, layer 24 |
|---|---|---|---|---|
| Anger | ↑ agh, -force, ough, force<br>↓ sill, fug, moment, trans | ↑ impot, fuck, throw, aim<br>↓ imposs, familiar, strange, secret | ↑ harsh, withdrawal, angry, dark<br>↓ smiles, joy, smiling, kins | ↑ angry, discipline, anger, harsh<br>↓ gai, smiling, pies, comed |
| Disgust | ↑ disposal, acob, arroj, -delete<br>↓ pend, tracing, cogn, stabil | ↑ vom, disgust, toilet, vomiting<br>↓ tumult, stolen, idle, peaceful | ↑ sour, stom, frown, bot<br>↓ Mat, hosp, ivan, ikel | ↑ sour, facial, frown, wrink<br>↓ tears, warm, Tears, light |
| Fear | ↑ whispers, uncertain, dread, constantly<br>↓ gen, va, rejo, otta | ↑ cla, nervously, frightened, dread<br>↓ smirk, POSE, aterra, cheer | ↑ gye, powered, gou, Cold<br>↓ leisure, pol, Pat, travel | ↑ horror, shock, shocked, Horror<br>↓ warm, maturity, boys, cke |
| Happiness | ↑ misch, joy, sill, light<br>↓ cold, inic, chos, .define | ↑ misch, joy, sunshine, infectious<br>↓ cold, cla, flesh, twice | ↑ smiles, joy, bright, smiling<br>↓ cold, Cold, kog, gou | ↑ smile, smiles, smiling, warm<br>↓ alarm, dramat, dram, distress |
| Sadness | ↑ wist, traced, Muller, stubborn<br>↓ anova, klas, !*, Terr | ↑ lonely, loneliness, triste, ghosts<br>↓ anova, throw, gotta, ailable | ↑ sorrow, tears, pain, experienced<br>↓ ris, grin, fun, incor | ↑ tears, Tears, crying, tear<br>↓ grin, teeth, tooth, smile |
| Surprise | ↑ sur, parc, Sur, Tent<br>↓ odi, emport, Abschluss, davon | ↑ suddenly, Suddenly, drop, sud<br>↓ abra, spite, davon, vind | ↑ rio, richtungen, paj, dada<br>↓ den, conditions, Bad, down | ↑ dada, nursery, lub, ris<br>↓ emotional, tears, frown, facial |

Figure 9: Top-4 promoted (↑) and suppressed (↓) tokens for story- and AffectNet-derived emotion vectors at layers 18 and 24. Scores use each model's final normalization and unembedding.

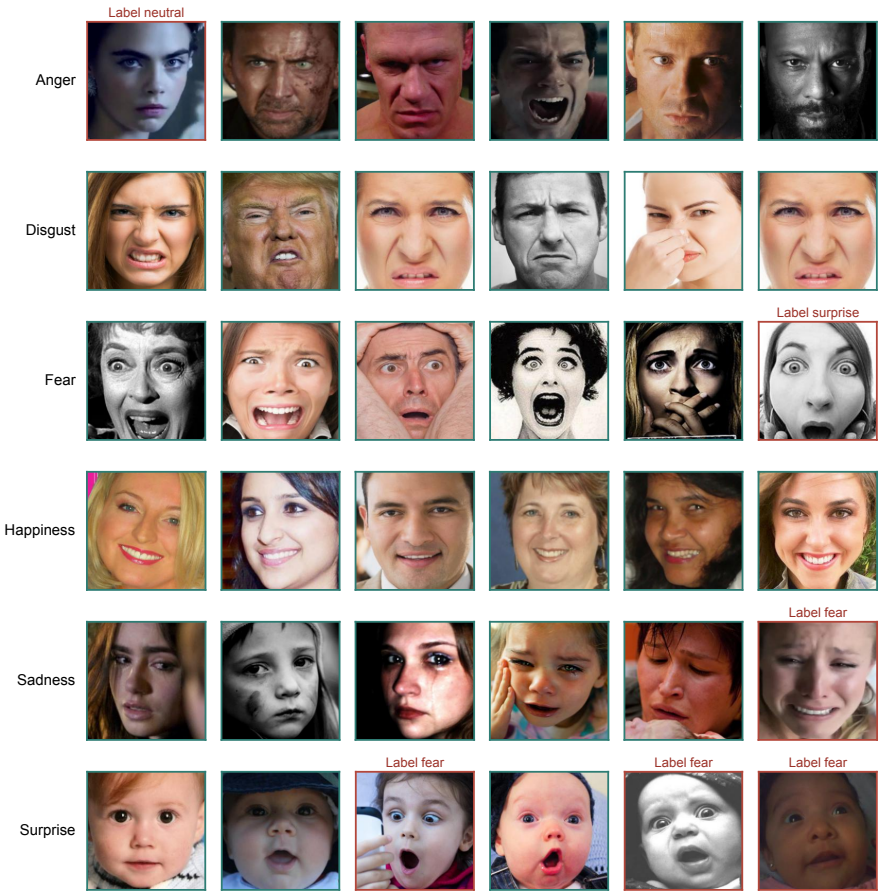



Figure 10: Six random images from each Qwen3-VL emotion vector's retrievals. Green frames mark label matches; red frames mark mismatches and show the ground-truth label.

example, Qwen3-VL's anger vector promotes tokens referring to a suit and family roles, which do not directly name the emotion.

At layer 24, the words promoted by image vectors are more closely related to emotion. The lowest-scoring tokens also often describe contrasting emotional states. For example, Qwen3-VL's AffectNet anger vector promotes *brutal* while suppressing *happy* and *joy*. These examples suggest that image vectors become easier to interpret in emotional terms at the later layer. However, the readouts are not exact emotion labels, and some tokens still describe facial appearance or related emotions. The comparison supports a change between the two sampled layers, similar to the depth-related pattern reported for text representations by Sofroniew et al. (2026).

### C.2 Qualitative Evidence from Retrieved Images and Patch-Level Readouts

**Images retrieved by each vector provide visual evidence of its meaning.** We next test whether an emotion vector retrieves images of the corresponding emotion. Figure 10 shows Qwen3-VL results on AffectNet validation images. We rank all 3,999 images by cosine similarity between their pooled activations and each vector, then randomly sample six examples from its top-50. The displayed disgust and happiness examples mostly match their target labels. Some mismatches involve related expressions, including fear and surprise in both directions. This is consistent with the overlap between these expressions reported in human facial perception (Ekman & Friesen, 1971; Smith et al., 2005).

Table 2: Retrieval P@50 and six-way classification recall for AffectNet-trained vectors. P@50 evaluates the top-50 retrievals; recall evaluates all validation images within each of the six target classes. All values are percentages.

| Emotion | Retrieval P@50 (%) | | | Classification Recall (%) | | |
|---|---|---|---|---|---|---|
| | Qwen3-VL | InternVL3.5 | Ministral-3 | Qwen3-VL | InternVL3.5 | Ministral-3 |
| anger | 58 | 60 | 38 | 41 | 52 | 29 |
| disgust | 80 | 66 | 68 | 34 | 37 | 30 |
| fear | 72 | 70 | 72 | 52 | 49 | 45 |
| happiness | 92 | 84 | 88 | 93 | 89 | 93 |
| sadness | 82 | 88 | 90 | 54 | 47 | 50 |
| surprise | 58 | 54 | 56 | 34 | 36 | 31 |

**The full top-50 sets support the retrieval pattern in the examples.** To check the pattern beyond the six displayed images, Table 2 reports precision at 50 (P@50), the fraction of target-class images among the top-50 retrievals for each emotion and model. The AffectNet pool contains eight approximately balanced classes, giving a reference rate of about 12.5%. Every emotion vector retrieves its target class well above this rate. Across models, happiness (84–92%) and sadness (82–90%) have the highest P@50. Disgust (66–80%) and fear (70–72%) are intermediate, while anger (38–60%) and surprise (54–58%) are generally lower. These results show that emotion vectors retrieve clear examples of their target classes. To explore whether the vectors also distinguish emotions beyond these top-ranked examples, we next evaluate six-way classification on all validation images from the six target classes.

For this six-way classification, we subtract the equally weighted mean of the six training-class activations from each pooled image activation, score it by its dot product with each unit emotion vector, and predict the highest-scoring class. We evaluate all validation images from the six target classes, excluding neutral and contempt where present. For each emotion, recall is the fraction of all images with that ground-truth label that the classifier predicts correctly. Across the 18 model–emotion combinations, P@50 and recall have a Spearman correlation of $\rho = +0.72$. Happiness performs well under both measures, but a clean top-ranked set does not always imply high class-wide recall. For example, Qwen3-VL's disgust vector has P@50 of 80% but recall of only 34%. Retrieval ranks images separately for each vector using uncentered cosine, whereas classification compares the six scores after centering. An emotion vector can therefore retrieve clear examples of its target class even when classification across the full dataset remains limited. For Qwen3-VL on AffectNet, vector-based classification performs at a broadly similar level to direct questioning. On the same validation images, accuracy is 51.4% for six-way vector classification and 47.6% for direct questioning.

**RAF-DB disgust vectors have low retrieval P@50 but retain classification ability.** We repeat the analysis with RAF-DB-trained vectors and the RAF-DB test set in Table 3. Because this pool is imbalanced, each P@50 value should be compared with the corresponding class's share of the pool. Five emotions show clear enrichment over their reference rates. Disgust differs, with P@50 of 8/16/6% for Qwen3-VL, InternVL3.5, and Ministral-3, compared with a pool share of 5.2%. The increase is small for Qwen3-VL and Ministral-3 and larger for InternVL3.5, but all three have much lower P@50 for disgust than on AffectNet. Anger has a similar pool share of 5.3% yet reaches P@50 of 72/66/64%, so class rarity alone does not explain the weak disgust retrievals.

Despite this low P@50, the centered six-way classifier has disgust recall of 60/48/53%, compared with 34/37/30% on AffectNet. These values come from different evaluation datasets, so they do not provide a controlled comparison of classification difficulty. They do show that weak disgust retrieval does not imply that the RAF-DB vectors cannot distinguish disgust from the other five classes. To understand what the vector retrieves instead, we next explore the labels within its top-50 set and compare this pattern with the vector geometry.

**Retrieval labels help explain the source-specific differences.** Table 4 breaks down the top-50 retrievals for disgust and fear by their ground-truth labels. On AffectNet, most mismatches for disgust are anger images, while most mismatches for fear are surprise images. On RAF-DB, anger

Table 3: Retrieval P@50 and six-way classification recall for RAF-DB-trained vectors. P@50 evaluates the top-50 retrievals; recall uses all test images within each target class. Pool share gives each class's reference rate in the 3,068-image retrieval pool. All values are percentages.

| Emotion | Pool share (%) | Retrieval P@50 (%) | | | Classification Recall (%) | | |
|---|---|---|---|---|---|---|---|
| | | Qwen3-VL | InternVL3.5 | Ministral-3 | Qwen3-VL | InternVL3.5 | Ministral-3 |
| anger | 5.3 | 72 | 66 | 64 | 69 | 51 | 54 |
| disgust | 5.2 | 8 | 16 | 6 | 60 | 48 | 53 |
| fear | 2.4 | 28 | 64 | 48 | 62 | 62 | 64 |
| happiness | 38.6 | 100 | 94 | 98 | 60 | 63 | 65 |
| sadness | 15.6 | 86 | 80 | 96 | 66 | 61 | 61 |
| surprise | 10.7 | 48 | 40 | 34 | 40 | 45 | 39 |

Table 4: Ground-truth composition (%) of the top-50 images retrieved by the disgust and fear vectors on AffectNet and RAF-DB. Bold values mark the vector's target class and equal its retrieval P@50. A dash denotes $< 0.5\%$.

| Emotion vector | Model | Disgust | Fear | Anger | Happiness | Neutral | Surprise | Sadness |
|---|---|---|---|---|---|---|---|---|
| *AffectNet validation pool* | | | | | | | | |
| disgust | Qwen3-VL | **80** | 2 | 18 | – | – | – | – |
| | InternVL3.5 | **66** | 4 | 26 | – | – | – | 4 |
| | Ministral-3 | **68** | 6 | 18 | – | – | – | 8 |
| fear | Qwen3-VL | 2 | **72** | 8 | – | – | 18 | – |
| | InternVL3.5 | 2 | **70** | 6 | – | – | 22 | – |
| | Ministral-3 | – | **72** | 6 | – | – | 22 | – |
| *RAF-DB test pool* | | | | | | | | |
| disgust | Qwen3-VL | **8** | – | 2 | 46 | 44 | – | – |
| | InternVL3.5 | **16** | – | 2 | 32 | 34 | 4 | 12 |
| | Ministral-3 | **6** | – | – | 40 | 48 | – | 6 |
| fear | Qwen3-VL | – | **28** | 24 | 10 | – | 32 | 6 |
| | InternVL3.5 | 2 | **64** | 22 | 2 | – | 8 | 2 |
| | Ministral-3 | 4 | **48** | 20 | 4 | – | 22 | 2 |

accounts for at most $2\%$ of the disgust retrievals. Happiness ($46/32/40\%$) and neutral ($44/34/48\%$) instead make up most of the set. Anger also accounts for a larger share of fear retrievals on RAF-DB ($24/22/20\%$) than on AffectNet ($8/6/6\%$), although fear and surprise remain prominent in these sets.

The vector geometry shows a related source difference, as shown in Figure 16 in Appendix D.1. Across the three models, RAF-DB places disgust closer to happiness and fear closer to anger, while the other sources more often associate disgust with anger and fear with surprise. Retrieval and vector geometry therefore provide complementary evidence. The first explores which labeled images score highly along each vector, while the second compares the six vectors directly. Together, they suggest that RAF-DB changes the relationships among some emotions, rather than simply weakening all emotion information.

**Patch-level readouts often highlight expressive facial regions.** We then explore where high-similarity image tokens occur. Figure 11 compares each image-token activation $\boldsymbol{h}_t$ directly with the corresponding emotion vector $\boldsymbol{v}_e$, before pooling the tokens. In the displayed examples, high-cosine regions often overlap facial areas involved in the expression. These include the brow and central face for anger, the nose and mouth for disgust, and the eyes and mouth for fear and surprise. Happiness examples often highlight the mouth area, while sadness examples include regions around the eyes and brow. Hair, clothing, and image borders are generally less prominent in these examples.

These maps provide a qualitative view of token similarity, not a test of which image regions cause the model's prediction. The images are tightly cropped around faces, so they also do not establish a preference for faces over natural backgrounds. The signal is spatially broad. A token's position

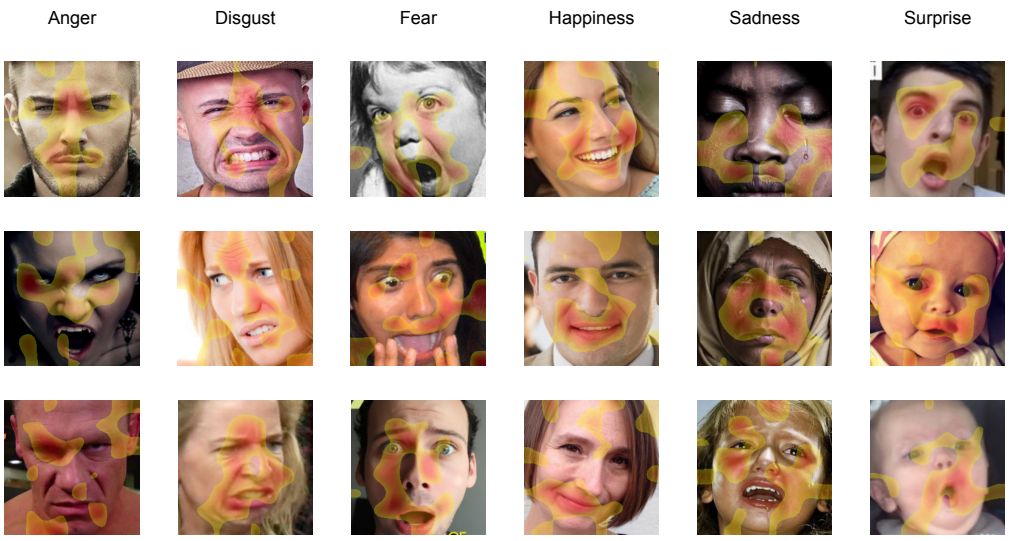



Figure 11: Patch cosine similarity to Qwen3-VL emotion vectors. Each emotion shows three label-matching AffectNet validation images sampled from its top-50 retrievals. Overlays show the top 30% of image tokens ranked by $\cos(\boldsymbol{h}_t, \boldsymbol{v}_e)$, with colors normalized for each image.

indicates where its visual input came from, but its later-layer activation can also contain information from other positions. The image-level retrieval result should therefore be understood as a property of the pooled visual representation, without assuming that one local patch carries the emotion signal.

### C.3 Emotion-Vector Structure Across Depth

We next explore whether the relationships among the six emotion vectors change across layers. At each layer, we compute the $6 \times 6$ emotion–emotion cosine matrix and keep its 15 unique off-diagonal entries. We then compare two layers by the cosine similarity between these 15-value lists. This compares the relationships among emotions, rather than the directions of individual vectors. Figure 12 shows the resulting layer-by-layer comparisons for T2 stories and AffectNet images.

**T2 geometry is stable across depth, while AffectNet geometry changes between layer ranges.** For T2 stories, the similarity between the emotion relationships is at least 0.89 for every pair of sampled layers in all three models. This is consistent with the text-based depth analysis of Sofroniew et al. (2026). AffectNet instead shows blocks of high similarity separated by changes across layers. Thus, its six-emotion geometry is relatively stable within some layer ranges but less stable across those ranges. Together with the lexical readouts in Appendix C.1, this suggests that the image-derived representation changes as the model processes the image.

### C.4 Principal Axes Across Models and Stimulus Sources

We next use PCA to identify the main dimensions of variation among the six vectors. Table 5 reports how much variance each principal component explains. It contains the 21 original model–source sets, six filtered image sets, and three native-frame copies of the same consensus set. PC1 explains 30–46% of the variance, and PC1+PC2 explain 55–75%, compared with 40% if variance were evenly spread across five dimensions. The first two components therefore capture more than half of the variance in every reported set and at least 60% in 25 of the 30 rows.

**The leading axes agree more consistently with valence than with arousal.** We compare the six emotions' PC1 coordinates with human valence and their PC2 coordinates with human arousal using the R&M (Russell & Mehrabian, 1977) and NRC-VAD (Mohammad, 2018) references. Figure 13 shows the correlations with human references and between emotion-vector sets. Among the 21

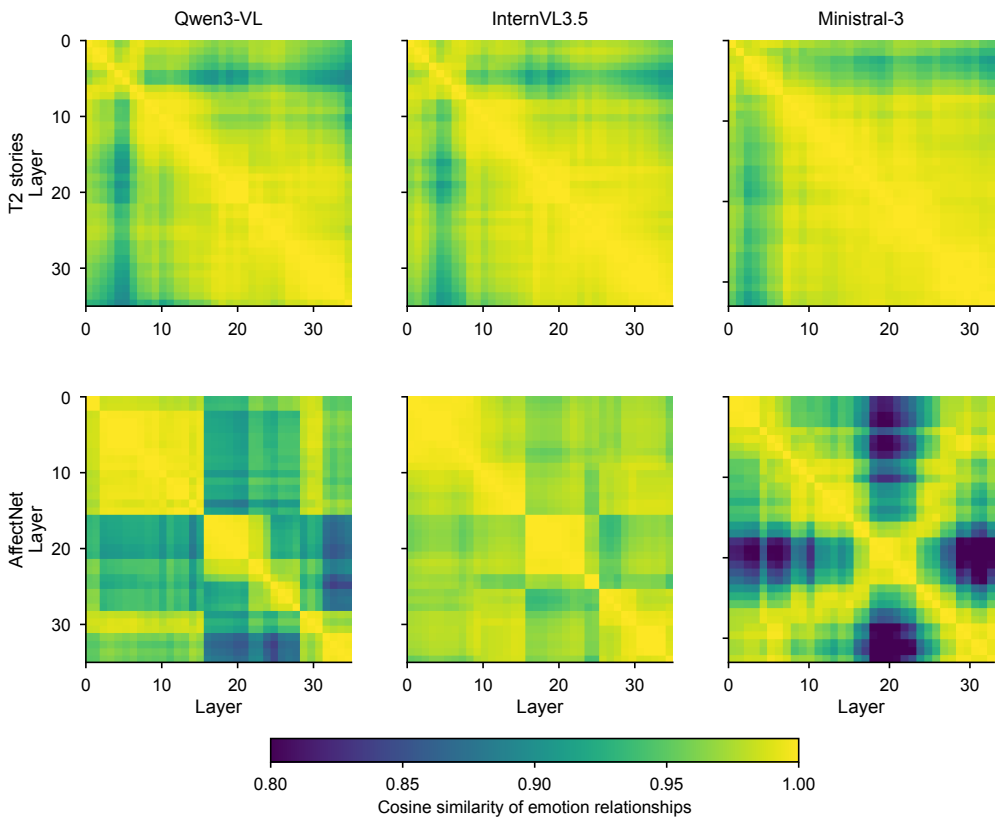



Figure 12: Cross-layer geometry of T2 story and AffectNet emotion vectors. Each cell compares two layers by the cosine similarity between the 15 unique off-diagonal entries of their emotion–emotion cosine matrices.

original model–source sets, PC1 has a median correlation of $+0.90$ with R&M valence, ranging from $+0.47$ to $+0.98$. The corresponding NRC-VAD median is $+0.93$, with a range of $+0.51$ to $+0.99$. Pairwise PC1 agreement among these sets also remains high, with a median of $+0.78$ over 210 pairs.

PC2 agreement with human arousal varies more. Its median correlation is $+0.58$ with R&M and $+0.66$ with NRC-VAD, while median pairwise PC2 agreement is $+0.41$. The lowest pairwise correlation is $-0.92$, between the Qwen3-VL and InternVL3.5 AffectNet sets. However, their correlations with R&M arousal are only $0.23$ and $0.03$, respectively. The negative pairwise value therefore describes their sign-oriented PC2 coordinates and should not be read as evidence that the models encode opposite human arousal orderings. More generally, each correlation uses only six emotions. As a descriptive comparison, 14 of the 21 sets exceed $0.8$ for valence against at least one human reference, compared with four for arousal. We use this cutoff to summarize the strength of agreement, not as a statistical significance threshold.

### C.5 Set-Level Similarity Across Stimulus Sources

The first two components leave about 25–45% of the variance outside the displayed plane. We therefore compare all six vectors in the full representation space. Figure 14 reports three measures, defined in Appendix B.7. Linear CKA compares the centered inner-product structure and is unchanged by an orthogonal rotation or a uniform change in scale. RSA compares the rank order of the 15 pairwise cosine distances between emotions. Cosine similarity compares each emotion vector directly with its counterpart in the native model coordinates. These measures answer different questions, so high agreement under one does not require high agreement under the others.

Table 5: PCA variance explained (%) for six centered emotion vectors. PC1+PC2 gives the cumulative share of the leading plane. Consensus rows share the same spectrum because orthogonal mapping into each model preserves variance shares (Appendix F).

| Stimulus source | PC1 | PC2 | PC3 | PC4 | PC5 | PC1+PC2 |
|---|---|---|---|---|---|---|
| *Qwen3-VL* | | | | | | |
| T1 | 30.7 | 25.2 | 19.2 | 13.9 | 11.1 | 55.9 |
| T2 | 38.2 | 27.6 | 20.3 | 8.4 | 5.5 | 65.8 |
| T3 | 40.2 | 25.9 | 19.9 | 8.0 | 6.1 | 66.0 |
| AffectNet | 34.0 | 29.8 | 18.8 | 11.6 | 5.8 | 63.8 |
| AffectNet (filtered) | 32.8 | 23.0 | 21.4 | 15.2 | 7.5 | 55.9 |
| RAF-DB | 43.0 | 19.9 | 17.5 | 12.6 | 6.9 | 63.0 |
| RAF-DB (filtered) | 38.5 | 24.5 | 17.1 | 11.9 | 8.1 | 62.9 |
| Synth-P | 38.7 | 29.9 | 22.5 | 6.4 | 2.5 | 68.6 |
| Synth-S | 42.6 | 24.3 | 16.6 | 10.7 | 5.8 | 66.9 |
| Consensus | 41.4 | 25.6 | 20.2 | 7.5 | 5.3 | 67.0 |
| *InternVL3.5* | | | | | | |
| T1 | 30.4 | 25.1 | 18.8 | 14.4 | 11.3 | 55.5 |
| T2 | 37.7 | 27.4 | 20.9 | 8.1 | 5.9 | 65.1 |
| T3 | 40.6 | 25.9 | 19.1 | 7.7 | 6.7 | 66.5 |
| AffectNet | 37.0 | 29.2 | 18.9 | 10.3 | 4.7 | 66.2 |
| AffectNet (filtered) | 31.9 | 26.1 | 21.3 | 14.5 | 6.2 | 58.0 |
| RAF-DB | 43.2 | 19.9 | 16.0 | 12.3 | 8.6 | 63.1 |
| RAF-DB (filtered) | 42.5 | 22.2 | 15.1 | 12.7 | 7.5 | 64.7 |
| Synth-P | 37.9 | 29.6 | 21.4 | 7.4 | 3.8 | 67.5 |
| Synth-S | 45.9 | 23.8 | 14.6 | 9.6 | 6.2 | 69.7 |
| Consensus | 41.4 | 25.6 | 20.2 | 7.5 | 5.3 | 67.0 |
| *Ministral-3* | | | | | | |
| T1 | 30.0 | 24.8 | 20.3 | 14.2 | 10.7 | 54.8 |
| T2 | 36.9 | 26.6 | 21.3 | 8.8 | 6.4 | 63.5 |
| T3 | 41.6 | 26.3 | 18.6 | 7.6 | 5.9 | 67.9 |
| AffectNet | 39.5 | 27.4 | 14.3 | 13.2 | 5.6 | 66.9 |
| AffectNet (filtered) | 37.4 | 23.1 | 17.2 | 14.5 | 7.8 | 60.5 |
| RAF-DB | 43.9 | 22.5 | 13.8 | 10.9 | 8.9 | 66.4 |
| RAF-DB (filtered) | 38.9 | 21.7 | 17.4 | 13.3 | 8.6 | 60.7 |
| Synth-P | 46.3 | 23.6 | 19.7 | 6.0 | 4.4 | 69.9 |
| Synth-S | 45.2 | 29.6 | 12.4 | 7.6 | 5.2 | 74.8 |
| Consensus | 41.4 | 25.6 | 20.2 | 7.5 | 5.3 | 67.0 |

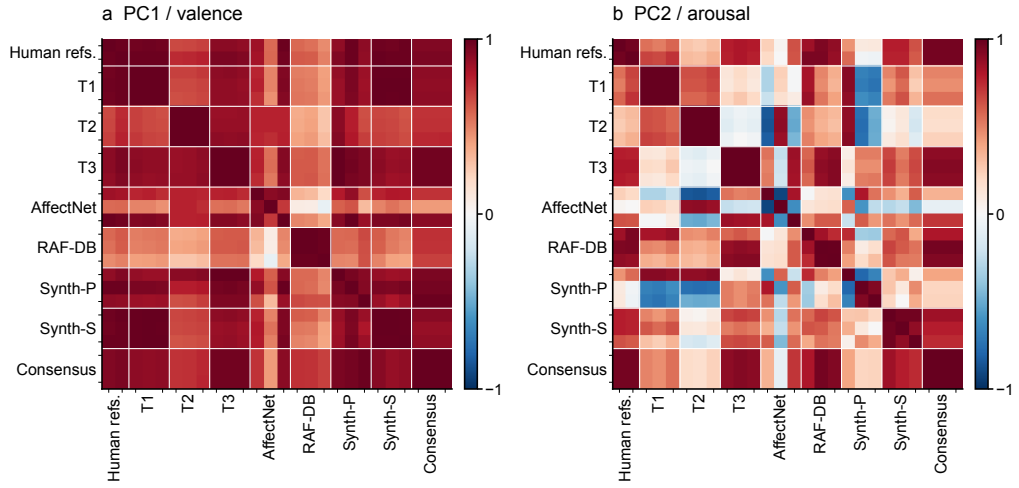



Figure 13: Pearson correlations across six-emotion PC1 (a, valence) and PC2 (b, arousal) coordinates, with signs aligned to R&M. Models are ordered Qwen3-VL, InternVL3.5, and Ministral-3 within each source and consensus group. Human references are ordered R&M, then NRC-VAD. Consensus vectors are mapped into each model's native frame.

**Changing stimulus source produces larger geometry differences than changing architecture.** To separate source and architecture effects, we compare pairs that differ in only one factor. We first hold the source fixed and vary the architecture. Across these 21 pairs, linear CKA ranges from 0.94 to 1.00, with a median of 0.99. The high agreement also holds for pairs involving Ministral-3, whose native coordinates differ strongly from those of Qwen3-VL and InternVL3.5. Thus, similar centered geometry does not depend on the models sharing native vector directions.

We then hold the architecture fixed and vary the stimulus source. Across these 63 pairs, CKA ranges from 0.69 to 0.98, with a median of 0.90 (Figure 15). Some sources remain similar, so the two distributions overlap. However, source changes can reduce CKA to 0.69, while architecture changes within a source keep it at or above 0.94. Within the models and sources studied here, source differences therefore have a larger effect on the centered six-emotion geometry.

**RSA makes the RAF-DB source difference more visible.** RSA shows the same source-versus-architecture pattern. Its median is 0.95 when the source is fixed and 0.66 when the architecture is fixed, compared with CKA medians of 0.99 and 0.90. RAF-DB accounts for much of the lower source agreement. Within each model, its median RSA with the other sources is 0.04, compared with 0.75 among the remaining sources. All 18 source pairs below 0.4 involve RAF-DB. In contrast, RAF-DB vectors from different architectures have RSA of 0.87–0.93. The RAF-DB pattern therefore repeats across all three models, although these comparisons alone do not rule out effects of the shared stimulus sample.

CKA also places RAF-DB apart from the other sources, but the difference is smaller. This suggests that RAF-DB changes the ranking of emotion relationships more strongly than their overall centered inner-product structure. The pattern is not a general text–image difference, because RAF-DB differs from both text sources and other image sources. Appendix D.1 explores the emotion pairs in more detail and identifies disgust and fear as the main contributors.

**Cosine similarity shows differences in native vector directions.** An orthogonal change of coordinates leaves CKA and RSA unchanged, but it can change cosine similarity between models. For Ministral-3 versus Qwen3-VL or InternVL3.5, matched-emotion cosine is near zero, as explored further in Appendix D.3. Qwen3-VL and InternVL3.5 instead have substantially higher native correspondence. Same-source text vectors reach approximately 0.93–0.94, while image vectors reach

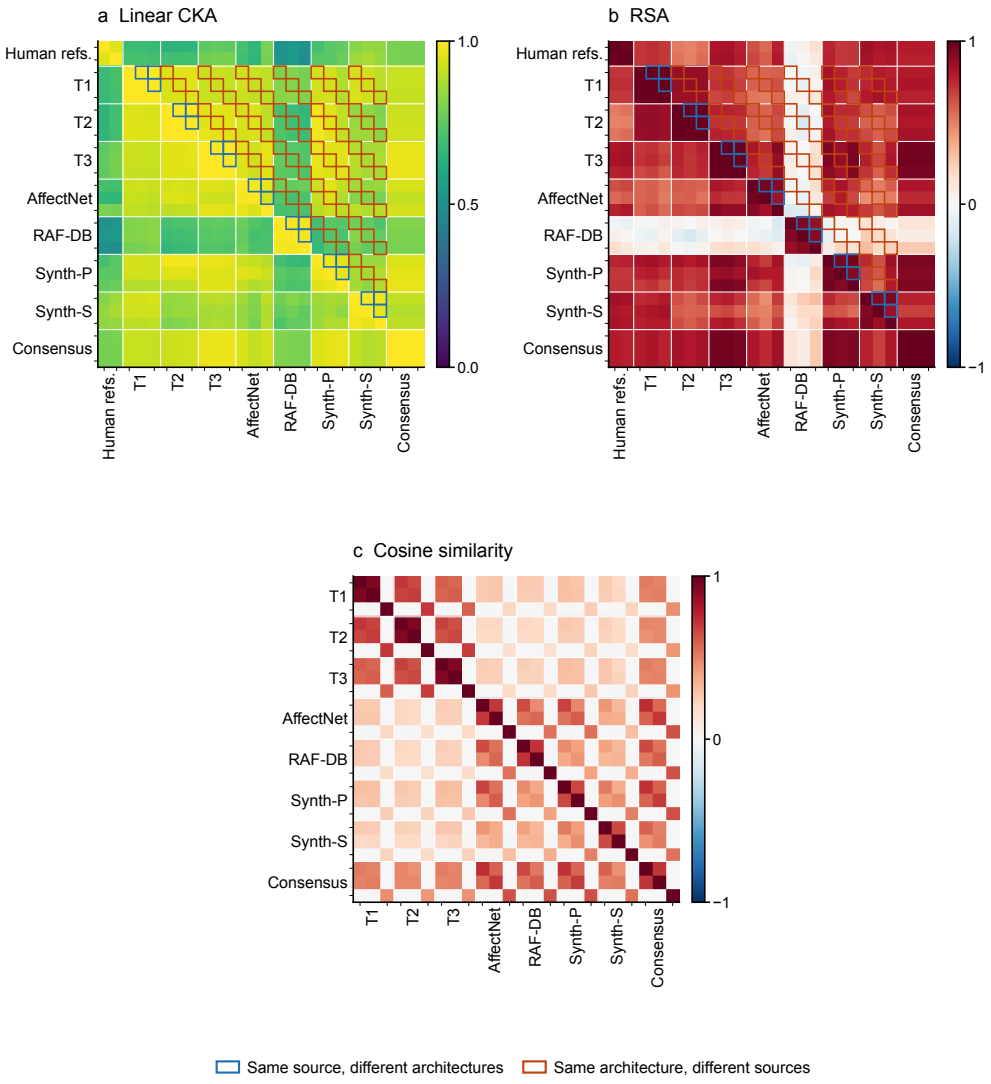



Figure 14: Pairwise linear CKA (a), RSA (b), and mean matched-emotion cosine similarity (c). Models are ordered Qwen3-VL, InternVL3.5, and Ministral-3 within each source and consensus group. Human references are ordered R&M, then NRC-VAD, and are omitted from (c) because they use different coordinates. Consensus vectors are mapped into each model's native frame. Blue outlines mark pairs with the same source and different architectures; orange outlines mark pairs with the same architecture and different sources.

0.66–0.72. InternVL3.5 uses a Qwen3-based language backbone, which may help explain this pattern.

**Full-set similarity and axis agreement provide related but different views.** We next compare these measures across the 210 pairs of original model–source sets. CKA and RSA are strongly associated with pairwise PC1 agreement, with Spearman correlations of 0.75 and 0.79. Their correlations with PC2 agreement are smaller, at 0.13 and 0.20. Cosine similarity has little association with either axis, with correlations of 0.03 for PC1 and $-0.03$ for PC2. These are descriptive comparisons among overlapping pairs, not independent observations. They support the distinction between the geometry of the six-emotion set and the native directions of its vectors.

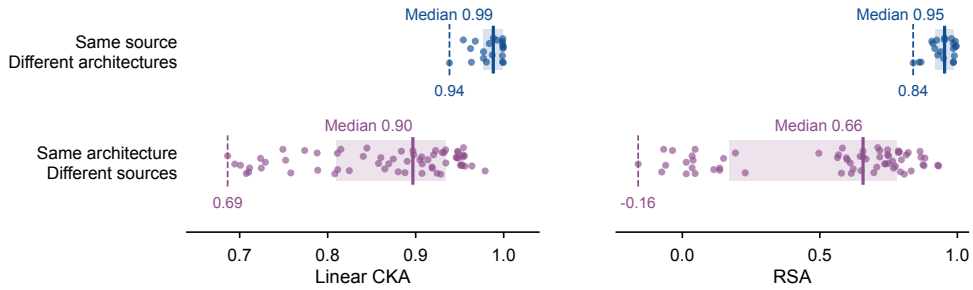



Figure 15: CKA and RSA for the 21 original model–source sets. Blue varies architecture with source fixed (21 pairs); purple varies source with architecture fixed (63 pairs). Points show all pairs and boxes show the interquartile range. Solid lines mark medians and dashed lines mark minima.

Table 6: Absolute Pearson correlations with R&M ratings across six emotions. PC1 is compared with valence and PC2 with arousal. Mean averages the three models; Spread is their maximum minus minimum. Sources are ordered by mean arousal agreement.

| Stimulus source | Qwen3-VL | InternVL3.5 | Ministral-3 | Mean | Spread |
|---|---|---|---|---|---|
| *Valence agreement* $\lvert r(\mathrm{PC1}, \mathrm{valence})\rvert$ | | | | | |
| RAF-DB | 0.57 | 0.55 | 0.47 | 0.53 | 0.10 |
| T3 | 0.90 | 0.91 | 0.88 | 0.90 | 0.02 |
| Synth-S | 0.97 | 0.97 | 0.98 | 0.97 | 0.01 |
| T1 | 0.96 | 0.97 | 0.96 | 0.97 | 0.01 |
| AffectNet | 0.85 | 0.56 | 0.98 | 0.80 | 0.42 |
| T2 | 0.66 | 0.67 | 0.68 | 0.67 | 0.02 |
| Synth-P | 0.88 | 0.98 | 0.90 | 0.92 | 0.10 |
| *Arousal agreement* $\lvert r(\mathrm{PC2}, \mathrm{arousal})\rvert$ | | | | | |
| RAF-DB | 0.85 | 0.94 | 0.89 | 0.89 | 0.09 |
| T3 | 0.78 | 0.81 | 0.79 | 0.80 | 0.03 |
| Synth-S | 0.78 | 0.79 | 0.64 | 0.73 | 0.15 |
| T1 | 0.55 | 0.51 | 0.58 | 0.55 | 0.07 |
| AffectNet | 0.23 | 0.03 | 0.64 | 0.30 | 0.61 |
| T2 | 0.27 | 0.25 | 0.30 | 0.28 | 0.05 |
| Synth-P | 0.45 | 0.10 | 0.11 | 0.22 | 0.35 |

The full-space measures also distinguish source pairs involving RAF-DB from the other source pairs. Both CKA and RSA give AUC values that round to 1.00, compared with 0.97 for PC1 agreement, while PC2 agreement provides little separation. Here, AUC summarizes how well each similarity score separates the two groups of source pairs. The small difference from PC1 alone does not establish that the full-space measures capture information independent of valence. Rather, axis agreement describes one dimension at a time, while CKA and RSA compare all relationships among the six vectors.

### C.6 Source Dependence of the Arousal Axis

The preceding sections show that PC1 generally agrees with human valence, while PC2 agreement with arousal varies more. We now explore whether this variation mainly follows stimulus source or model architecture. We then test whether selecting stimuli that a model recognizes correctly reduces the variation, and whether changes in individual axes also change the consensus vectors.

**Arousal agreement varies widely across stimulus sources.** Table 6 reports each model's agreement with human ratings and summarizes the mean and spread within each source. Averaged across the three models, arousal agreement ranges from 0.22 for synthetic portraits to 0.89 for RAF-DB. This between-source range is 0.67, while the average within-source range across models is 0.19.

Table 7: Effect of success filtering on the two leading emotion axes, measured as sign-aligned $|r|$ against the R&M norms. The same subset selected by Qwen3-VL is used for all three architectures. Retained sample sizes are reported in Table 1.

| | | Arousal $\|r\|$ (PC2) | | Valence $\|r\|$ (PC1) | |
|---|---|---|---|---|---|
| Source | Model | Unfiltered | Filtered | Unfiltered | Filtered |
| AffectNet | Qwen3-VL | 0.23 | 0.41 | 0.85 | 0.97 |
| | InternVL3.5 | 0.03 | 0.63 | 0.56 | 0.98 |
| | Ministral-3 | 0.64 | 0.88 | 0.98 | 0.97 |
| RAF-DB | Qwen3-VL | 0.85 | 0.65 | 0.57 | 0.67 |
| | InternVL3.5 | 0.94 | 0.79 | 0.55 | 0.65 |
| | Ministral-3 | 0.89 | 0.85 | 0.47 | 0.56 |
| T1 | Qwen3-VL | 0.55 | 0.71 | 0.96 | 0.93 |
| | InternVL3.5 | 0.51 | 0.70 | 0.97 | 0.90 |
| | Ministral-3 | 0.58 | 0.70 | 0.96 | 0.91 |
| T2 | Qwen3-VL | 0.27 | 0.77 | 0.66 | 0.95 |
| | InternVL3.5 | 0.25 | 0.83 | 0.67 | 0.93 |
| | Ministral-3 | 0.30 | 0.81 | 0.68 | 0.97 |
| T3 | Qwen3-VL | 0.78 | 0.59 | 0.90 | 0.75 |
| | InternVL3.5 | 0.81 | 0.62 | 0.91 | 0.78 |
| | Ministral-3 | 0.79 | 0.66 | 0.88 | 0.85 |

Valence agreement varies less, with source means from $0.53$ to $0.97$ and an average within-source model range of $0.10$. These summaries describe the observed variation, rather than a formal decomposition of source and model effects.

Both modalities contain sources with weak and strong arousal agreement. Among text sources, the means range from $0.28$ for T2 to $0.80$ for T3. Among image sources, they range from $0.22$ for synthetic portraits to $0.89$ for RAF-DB. The variation therefore does not follow a simple text–image division. Some sources also vary substantially across models. For example, AffectNet has arousal agreement from $0.03$ to $0.64$, so the larger overall source range does not mean that model differences are absent.

**Selecting correctly recognized stimuli can raise or lower axis agreement.** To test whether hard-to-recognize stimuli account for the source variation, we re-estimate emotion vectors using only stimuli that Qwen3-VL classifies correctly. Within each source, we balance the six emotion classes to the smallest retained class. All three models use the same selected examples. Thus, selection depends on Qwen3-VL's predictions, not on the predictions of InternVL3.5 or Ministral-3. We refer to this procedure as success filtering.

Table 7 shows that filtering changes arousal agreement in both directions. AffectNet rises from $0.03$–$0.64$ to $0.41$–$0.88$, while RAF-DB falls from $0.85$–$0.94$ to $0.65$–$0.85$. Text sources show the same contrast. T2 rises from $0.25$–$0.30$ to $0.77$–$0.83$, whereas T3 falls from $0.78$–$0.81$ to $0.59$–$0.66$. The correlation between the initial agreement and its change is $r = -0.91$ across these model–source settings. However, the change already includes subtraction of the initial value, so this correlation is not separate evidence that filtering corrects noise. The direct comparisons show that selecting correctly recognized stimuli changes the estimated axes without consistently moving them closer to the human reference.

**Size-matched subsets help separate sample-count and selection effects.** Filtering also reduces the sample count. As shown in Table 8, we therefore compare the filtered AffectNet set with random subsets of the same size, using 377 examples per class and five random seeds. The clearest selection effect occurs for InternVL3.5. Its arousal agreement is $0.03 \pm 0.03$ under random subsampling but $0.63$ after filtering. Ministral-3 shows a smaller difference, from $0.69 \pm 0.07$ to $0.88$. Qwen3-VL is less clear, with $0.37 \pm 0.23$ for random subsets and $0.41$ after filtering.

Table 8: AffectNet sample-size control. Random and filtered subsets contain 377 images per class. Random results are mean ± standard deviation over five seeds. Valence and arousal are correlations of PC1 and PC2 with human ratings. CKA and RSA compare image vectors with the same model's T2 text vectors.

| | R&M | | NRC-VAD | | vs. T2 text | |
|---|---|---|---|---|---|---|
| AffectNet subset | Valence | Arousal | Valence | Arousal | CKA | RSA |
| *Qwen3-VL* | | | | | | |
| Full set | 0.85 | 0.23 | 0.83 | 0.20 | 0.93 | 0.61 |
| Random subset | $0.81 \pm 0.22$ | $0.37 \pm 0.23$ | $0.79 \pm 0.22$ | $0.34 \pm 0.23$ | $0.92 \pm 0.02$ | $0.64 \pm 0.05$ |
| Success-filtered | 0.97 | 0.41 | 0.99 | 0.28 | 0.95 | 0.86 |
| *InternVL3.5* | | | | | | |
| Full set | 0.56 | 0.03 | 0.57 | 0.00 | 0.95 | 0.60 |
| Random subset | $0.56 \pm 0.16$ | $0.03 \pm 0.03$ | $0.57 \pm 0.15$ | $0.02 \pm 0.06$ | $0.94 \pm 0.01$ | $0.66 \pm 0.03$ |
| Success-filtered | 0.98 | 0.63 | 0.98 | 0.55 | 0.93 | 0.77 |
| *Ministral-3* | | | | | | |
| Full set | 0.98 | 0.64 | 0.97 | 0.63 | 0.91 | 0.66 |
| Random subset | $0.98 \pm 0.01$ | $0.69 \pm 0.07$ | $0.98 \pm 0.02$ | $0.68 \pm 0.08$ | $0.90 \pm 0.01$ | $0.63 \pm 0.05$ |
| Success-filtered | 0.97 | 0.88 | 0.98 | 0.85 | 0.90 | 0.65 |

**Arousal agreement and shared emotion structure can change differently.** RAF-DB has the highest mean arousal agreement in Table 6, at $0.89$, yet the weakest cross-source RSA in Appendix C.5. AffectNet shows the reverse pattern, with lower arousal agreement but stronger relational agreement with the other non-RAF-DB sources. Agreement with one human axis therefore does not summarize how well a source preserves the relationships among all six emotions.

We also check whether filtering changes the consensus vectors constructed in Appendix F. In this comparison, we replace only the AffectNet and RAF-DB vectors with their filtered versions, keeping the other five sources and the activation-based alignment setting unchanged. The two versions have mean matched-emotion cosine of $0.98$ and a minimum of about $0.95$. The planes formed by their first two principal components also remain close, with principal angles of $6°$–$18°$. These angles compare the leading two-dimensional planes, not the full shared subspaces. The result supports stability of the consensus vectors under this specific filtering comparison, even though individual source axes can change substantially. Our main analyses use the full labeled stimulus sets.

# D Representational Similarity

Appendix C shows that emotion vectors can share relationships even when their individual directions differ. We now explore whether vectors for the same emotion correspond across stimulus sources, modalities, and architectures. These comparisons complement the set-level geometry in Appendix C.5 and provide the basis for the causal-transfer experiments in Appendix E.

We first vary the stimulus source within the same model and modality in Appendix D.1. We then compare text and image vectors within each model in Appendix D.2, followed by comparisons across architectures in Appendix D.3. This order helps separate source differences from changes in modality and model coordinates.

We report four measures with different roles. Cosine similarity before alignment compares matched emotions in their original coordinates. Full-fit cosine uses all six emotion pairs to fit and evaluate an orthogonal transformation, so it measures fit to known pairs. LOEO cosine fits on five emotions and evaluates the sixth, testing whether the relationships learned from the other emotions predict a held-out pair. The permutation $z$ score compares mean LOEO cosine with the result of shuffling emotion labels and refitting. We use $z > 2$ as the reporting threshold. All cosine values are averaged over the six emotions unless stated otherwise. Appendix B.7 gives the definitions and controls.

Table 9: Within-modality correspondence. Each row changes the stimulus source within one model. Cosines are means over six emotions. Full-fit uses all six pairs. LOEO fits five and evaluates the sixth. The $z$ score compares LOEO with the refitted label-permutation null. Source labels follow Table 1.

| | | Mean cosine | | | Permutation |
|---|---|---|---|---|---|
| Model | Source pair | Before alignment | Full-fit | LOEO | $z$ |
| *Text sources* | | | | | |
| Qwen3-VL | T1–T2 | 0.71 | 0.98 | 0.72 | 4.09 |
| | T1–T3 | 0.61 | 0.96 | 0.60 | 4.38 |
| | T2–T3 | 0.69 | 0.98 | 0.72 | 4.52 |
| InternVL3.5 | T1–T2 | 0.70 | 0.98 | 0.71 | 4.26 |
| | T1–T3 | 0.60 | 0.96 | 0.60 | 4.42 |
| | T2–T3 | 0.67 | 0.98 | 0.70 | 4.57 |
| Ministral-3 | T1–T2 | 0.71 | 0.98 | 0.73 | 4.16 |
| | T1–T3 | 0.60 | 0.95 | 0.59 | 4.32 |
| | T2–T3 | 0.69 | 0.98 | 0.72 | 4.64 |
| *Image sources* | | | | | |
| Qwen3-VL | AffectNet–RAF-DB | 0.65 | 0.92 | 0.43 | 3.19 |
| | AffectNet–Synth-P | 0.68 | 0.96 | 0.65 | 4.18 |
| | AffectNet–Synth-S | 0.44 | 0.95 | 0.42 | 3.83 |
| | RAF-DB–Synth-P | 0.47 | 0.86 | 0.27 | 2.08 |
| | RAF-DB–Synth-S | 0.40 | 0.92 | 0.35 | 2.59 |
| | Synth-P–Synth-S | 0.52 | 0.93 | 0.50 | 3.75 |
| InternVL3.5 | AffectNet–RAF-DB | 0.57 | 0.90 | 0.36 | 3.02 |
| | AffectNet–Synth-P | 0.61 | 0.97 | 0.66 | 4.15 |
| | AffectNet–Synth-S | 0.36 | 0.94 | 0.39 | 3.37 |
| | RAF-DB–Synth-P | 0.44 | 0.87 | 0.28 | 2.33 |
| | RAF-DB–Synth-S | 0.34 | 0.93 | 0.33 | 2.72 |
| | Synth-P–Synth-S | 0.46 | 0.94 | 0.47 | 3.84 |
| Ministral-3 | AffectNet–RAF-DB | 0.55 | 0.89 | 0.37 | 2.77 |
| | AffectNet–Synth-P | 0.55 | 0.94 | 0.59 | 3.95 |
| | AffectNet–Synth-S | 0.30 | 0.95 | 0.46 | 3.65 |
| | RAF-DB–Synth-P | 0.40 | 0.85 | 0.30 | 2.88 |
| | RAF-DB–Synth-S | 0.36 | 0.89 | 0.33 | 2.75 |
| | Synth-P–Synth-S | 0.51 | 0.95 | 0.59 | 4.28 |

### D.1 Within-Modality Consistency Across Stimulus Sources

We first explore how stimulus source affects matched-emotion correspondence while keeping model and modality fixed. Table 9 compares vectors from different story generators and different image datasets under this protocol.

**Text vectors show more consistent correspondence across sources than image vectors.** Table 9 shows consistent correspondence across the three story sets. Cosine similarity before alignment ranges from 0.60 to 0.71, and LOEO cosine from 0.59 to 0.73. All comparisons exceed the permutation threshold ($z = 4.09$–$4.64$). Thus, the relationship learned from five emotions also predicts the sixth when the story generator changes.

Image-source comparisons vary more, with cosine similarity before alignment ranging from 0.30 to 0.68. However, the three models rank the six image-source pairs almost identically by cosine similarity (Spearman $\rho = 0.94$–$1.00$). AffectNet and synthetic portraits show strong held-out correspondence in every model ($z = 3.95$–$4.18$), while several pairs involving RAF-DB are weaker. The repeated ordering across models suggests that these differences depend strongly on the stimulus sources, consistent with the set-level comparison in Appendix C.5.

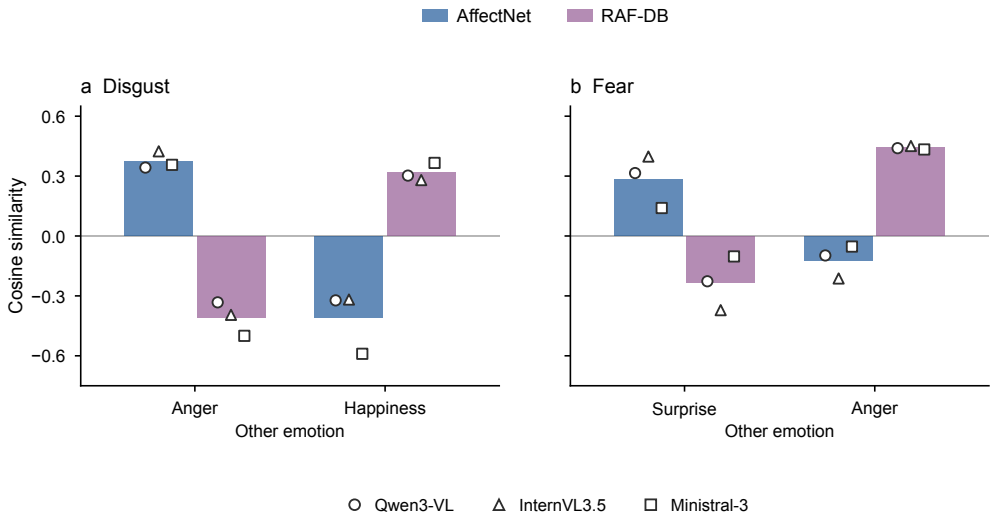



Figure 16: Source differences in the neighbors of disgust (a) and fear (b). Cosine similarities are computed between emotion vectors from AffectNet or RAF-DB. Bars show the mean over three models, and marker shapes identify individual models.

**Held-out alignment distinguishes generalization from fit to known pairs.** To determine whether high fitted similarity also predicts an unseen pair, we compare full-fit and LOEO cosine in Table 9. Full-fit cosine is high for all image-source pairs (0.85–0.97). Yet RAF-DB pairs give lower LOEO cosine (0.27–0.43) than the strongest comparisons without RAF-DB. All nine RAF-DB comparisons still exceed the permutation threshold ($z = 2.08$–$3.19$). Their correspondence is therefore weaker, but the other five emotions still predict the held-out pair better than shuffled labels do.

**RAF-DB changes the neighbors of disgust and fear.** To explore which emotion relationships differ, Figure 16 compares the neighbors of disgust and fear. Across the three models, AffectNet places disgust closer to anger and fear closer to surprise, whereas RAF-DB places disgust closer to happiness and fear closer to anger. The retrieval-label patterns in Table 4 provide related evidence. These results suggest that the two datasets emphasize different visual forms of the same labeled emotions, although they do not identify the visual features responsible for the shift.

Changes in a few emotion relationships can reduce RSA and held-out correspondence even when cosine similarity before alignment remains moderate. In a separate classification check, RAF-DB-derived probes classify 3,000 AffectNet images balanced across the six target classes. Accuracy is 0.38–0.42, above the uniform-chance reference of 0.167, while disgust and fear recall remain below each model's overall accuracy. These observations support a source-specific change in emotion relationships rather than a uniform loss of emotion information.

### D.2 Correspondence Between Text- and Image-Derived Vectors

We next explore whether text- and image-derived vectors correspond within the same model. For each architecture, Table 10 compares each of the three story sources with each of the four image sources. This changes both modality and source, so the within-modality results above provide a reference for interpreting the differences.

**Most source pairs retain held-out correspondence despite low cosine similarity.** Table 10 shows low cosine similarity before alignment across all 36 text–image comparisons (0.15–0.31). However, for AffectNet and the two synthetic image sources, LOEO cosine reaches 0.29–0.53, and all 27 comparisons exceed $z = 2$. Thus, the relationships among five emotions can predict the sixth across modalities even when matched text and image vectors have low similarity in their original coordinates.

Table 10: Text–image correspondence within each model. Cosines are means over six emotions. Full-fit uses all six pairs. LOEO fits five and evaluates the sixth. The $z$ score compares LOEO with the refitted label-permutation null. Null controls replace image vectors with matched-norm directions. Source labels follow Table 1.

| | Mean cosine | | | Permutation |
|---|---|---|---|---|
| Text → image source | Before alignment | Full-fit | LOEO | $z$ |
| *Qwen3-VL* | | | | |
| T1 → AffectNet | 0.26 | 0.97 | 0.33 | 4.34 |
| T1 → RAF-DB | 0.26 | 0.94 | 0.21 | 2.65 |
| T1 → Synth-P | 0.30 | 0.94 | 0.38 | 4.61 |
| T1 → Synth-S | 0.26 | 0.97 | 0.34 | 4.29 |
| T2 → AffectNet | 0.21 | 0.98 | 0.40 | 3.77 |
| T2 → RAF-DB | 0.20 | 0.90 | 0.12 | 1.07 |
| T2 → Synth-P | 0.27 | 0.98 | 0.53 | 4.21 |
| T2 → Synth-S | 0.24 | 0.94 | 0.29 | 2.62 |
| T3 → AffectNet | 0.25 | 0.98 | 0.43 | 4.34 |
| T3 → RAF-DB | 0.24 | 0.89 | 0.20 | 2.01 |
| T3 → Synth-P | 0.31 | 0.98 | 0.53 | 4.18 |
| T3 → Synth-S | 0.28 | 0.95 | 0.36 | 3.34 |
| *InternVL3.5* | | | | |
| T1 → AffectNet | 0.27 | 0.97 | 0.33 | 4.27 |
| T1 → RAF-DB | 0.25 | 0.94 | 0.20 | 2.77 |
| T1 → Synth-P | 0.28 | 0.95 | 0.33 | 4.36 |
| T1 → Synth-S | 0.21 | 0.97 | 0.33 | 4.14 |
| T2 → AffectNet | 0.20 | 0.98 | 0.43 | 3.70 |
| T2 → RAF-DB | 0.20 | 0.89 | 0.08 | 0.89 |
| T2 → Synth-P | 0.25 | 0.97 | 0.45 | 3.83 |
| T2 → Synth-S | 0.20 | 0.95 | 0.31 | 2.78 |
| T3 → AffectNet | 0.24 | 0.97 | 0.40 | 3.85 |
| T3 → RAF-DB | 0.23 | 0.89 | 0.16 | 1.83 |
| T3 → Synth-P | 0.29 | 0.98 | 0.44 | 3.75 |
| T3 → Synth-S | 0.22 | 0.95 | 0.34 | 3.15 |
| *Ministral-3* | | | | |
| T1 → AffectNet | 0.21 | 0.96 | 0.29 | 3.77 |
| T1 → RAF-DB | 0.19 | 0.94 | 0.20 | 2.73 |
| T1 → Synth-P | 0.25 | 0.91 | 0.34 | 4.48 |
| T1 → Synth-S | 0.23 | 0.95 | 0.33 | 4.22 |
| T2 → AffectNet | 0.16 | 0.97 | 0.36 | 3.49 |
| T2 → RAF-DB | 0.15 | 0.90 | 0.08 | 0.80 |
| T2 → Synth-P | 0.22 | 0.94 | 0.43 | 4.26 |
| T2 → Synth-S | 0.19 | 0.93 | 0.29 | 2.64 |
| T3 → AffectNet | 0.20 | 0.98 | 0.42 | 3.98 |
| T3 → RAF-DB | 0.19 | 0.89 | 0.17 | 1.81 |
| T3 → Synth-P | 0.27 | 0.98 | 0.50 | 4.06 |
| T3 → Synth-S | 0.23 | 0.95 | 0.36 | 3.04 |
| *Null controls for Qwen3-VL, T2 → AffectNet* | | | | |
| Random directions | −0.01 | 0.86 | 0.00 | −0.58 |
| Orthogonal directions | −0.01 | 0.86 | 0.00 | −0.45 |

Table 11: Ministral-3→Qwen3-VL correspondence. Cosines are means over six emotions. Before alignment, cosine similarity is measured in native coordinates. Full-fit and LOEO use six and five emotion pairs, respectively. RAF-DB controls replace source or target vectors with matched-norm directions. The $z$ score compares LOEO with the refitted label-permutation null. Source labels follow Table 1.

| | Mean cosine | | | Permutation |
|---|---|---|---|---|
| Source / control | Before alignment | Full-fit | LOEO | $z$ |
| T1 | 0.00 | 1.00 | 0.13 | 3.32 |
| T2 | 0.00 | 1.00 | 0.34 | 3.39 |
| T3 | 0.00 | 1.00 | 0.38 | 3.73 |
| AffectNet | 0.00 | 0.99 | 0.30 | 3.16 |
| RAF-DB | −0.01 | 0.99 | 0.23 | 2.60 |
| Synth-P | 0.00 | 0.98 | 0.45 | 3.63 |
| Synth-S | 0.00 | 0.99 | 0.36 | 3.28 |
| *RAF-DB null controls* | | | | |
| Random (target) | −0.01 | 0.87 | 0.00 | −0.58 |
| Random (source) | 0.01 | 0.87 | 0.02 | 1.92 |
| Orthogonal (target) | −0.01 | 0.87 | 0.00 | −0.58 |
| Orthogonal (source) | 0.01 | 0.87 | 0.02 | 1.92 |

The controls show why fitting and evaluating the same emotion pairs is insufficient. Full-fit cosine reaches 0.89–0.98 for the text–image pairs, but also reaches 0.86 after replacing image vectors with matched-norm random or orthogonal directions. These replacement controls instead give near-zero LOEO cosine. The held-out result and its label-permutation control therefore provide the evidence for emotion correspondence.

**Correspondence and causal transfer test different properties.** The LOEO result concerns relationships among emotion vectors. It does not establish whether a text-derived vector can act on image representations without alignment. Appendix E.2 tests that question directly by injecting text-derived vectors at image-token positions (Tables 16 and 17).

**RAF-DB correspondence is weaker within and across modalities.** To explore whether weaker text–image correspondence is specific to the modality change, we compare RAF-DB across both settings. Its text–image LOEO cosine is 0.08–0.21, with only four of nine comparisons above $z = 2$ ($z = 0.80$–$2.77$). RAF-DB also gives weaker image–image correspondence in Table 9, although all those comparisons exceed the threshold. The reduction therefore appears within as well as across modalities, consistent with the source differences in Appendix D.1.

**The same source ordering appears across architectures.** Averaging LOEO cosine over the three story sources gives the same broad ordering in every model. Synthetic portraits rank highest and RAF-DB lowest, with AffectNet and synthetic scenes between them. Together with the within-modality results, this suggests that stimulus source helps explain which emotion relationships transfer across text and images.

### D.3 Cross-Architecture Correspondence and Alignment

We further explore whether held-out emotion correspondence survives changes in model architecture. Table 11 compares Ministral-3 with Qwen3-VL, a pair with nearly zero native cosine, across all seven stimulus sources. We then test whether mappings learned from paired image activations can recover correspondence without fitting to the six emotion vectors.

**Cross-architecture correspondence survives near-zero cosine similarity.** Table 11 shows near-zero cosine similarity before alignment for every source. Nevertheless, LOEO cosine is positive for all seven sources (0.13–0.45), and all exceed the permutation threshold ($z = 2.60$–$3.73$). Random

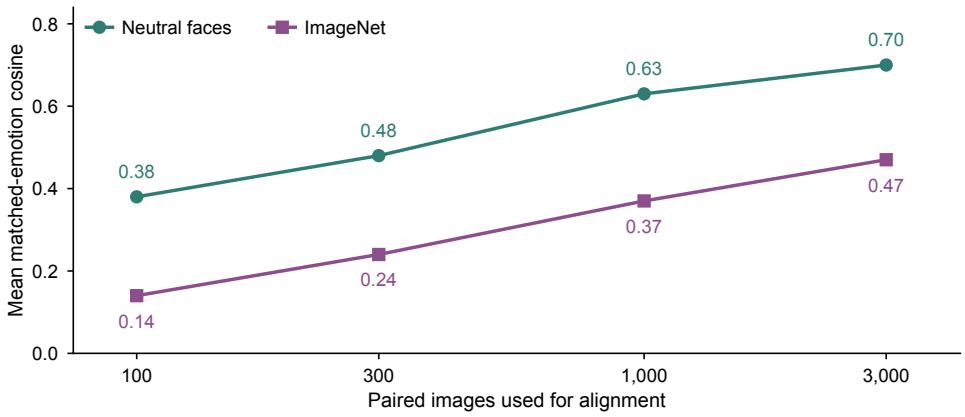



Figure 17: Alignment sample size and Ministral-3→Qwen3-VL correspondence. Maps use paired neutral-face or ImageNet activations without emotion-vector fitting. Points show mean cosine across six AffectNet emotion pairs. Sample size is on a logarithmic axis. Reference cosine similarities are 0.00 before alignment, 0.30 with LOEO, and 0.99 with a full fit.

and orthogonal replacement controls give near-zero held-out cosine. Thus, emotion relationships learned from five pairs can generalize to the sixth despite differences in model coordinates.

Full-fit cosine is again less informative because replacement controls already reach 0.87. We therefore use LOEO and its permutation control to assess correspondence beyond the emotion pairs used to fit the map.

**Learning a mapping without emotion-vector pairs.** The LOEO results in Tables 10 and 11 show that mappings learned from five matched emotion pairs can generalize to the sixth across modalities and architectures. This motivates us to explore whether we can learn a mapping without using emotion vectors or their labels and then apply it to all six emotion vectors. Following Section 5, we use activations from the same neutral-face or ImageNet images in both models to fit an orthogonal transformation (Equation 3). We vary the number of paired images used for fitting and apply each resulting mapping unchanged to all six AffectNet emotion vectors (Figure 17).

**Emotion correspondence improves with more paired images.** As shown in Figure 17, neutral-face mapping raises mean cosine from 0.38 with 100 paired images to 0.70 with 3,000. To test whether this recovery depends on images from an emotion dataset, we also fit the mapping with ImageNet. Mean cosine increases from 0.14 with 100 paired images to 0.47 with 3,000. Both mappings improve correspondence at every tested sample size, although neutral faces give higher cosine.

Together, these results suggest that a mapping learned independently of the six emotion vectors can recover their correspondence, even when the fitting images come from outside the emotion datasets. This supports the view that differences in model coordinates account for part of the low cosine similarity before alignment. We next test whether this recovered correspondence also supports causal transfer by injecting the mapped vectors into the target model (Table 18 in Appendix E.3).

**Held-out correspondence is distributed across emotions.** To check whether a single emotion drives the mean, Table 12 reports each held-out cosine. Several emotions contribute positive correspondence in every source. However, strength varies by emotion, and RAF-DB sadness and surprise are close to zero. The permutation result therefore supports correspondence for the six-emotion set as a whole, without implying equally strong correspondence for every emotion.

Table 12: Per-emotion LOEO cosine for Ministral-3→Qwen3-VL. Each cell evaluates the named emotion after fitting the other five. Source labels follow Table 1.

| Stimulus source | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| T1 | 0.10 | 0.11 | 0.21 | 0.15 | 0.12 | 0.07 |
| T2 | 0.50 | 0.54 | 0.37 | 0.25 | 0.17 | 0.20 |
| T3 | 0.53 | 0.55 | 0.44 | 0.31 | 0.21 | 0.24 |
| AffectNet | 0.26 | 0.28 | 0.46 | 0.31 | 0.21 | 0.31 |
| RAF-DB | 0.36 | 0.26 | 0.41 | 0.35 | 0.01 | 0.00 |
| Synth-P | 0.64 | 0.66 | 0.56 | 0.24 | 0.34 | 0.23 |
| Synth-S | 0.33 | 0.32 | 0.64 | 0.46 | 0.18 | 0.23 |

Table 13: LOEO after activation-based mapping from Ministral-3 to Qwen3-VL. Direct and After LOEO give mean cosine over six emotions. Held-out range gives the six LOEO scores. The $z$ score uses refitted label permutations. Each activation-based map uses 3,000 paired images. No mapping gives ordinary LOEO. A random map rotates the real source vectors, whereas the final control replaces them with target-orthogonal random directions.

| | Mean cosine | | | |
|---|---|---|---|---|
| Initial condition | Direct | After LOEO | Held-out range | Permutation $z$ |
| *Image vectors (AffectNet)* | | | | |
| No mapping | 0.000 | 0.305 | 0.206–0.456 | 3.16 |
| Neutral-face map | 0.697 | 0.755 | 0.694–0.827 | 4.37 |
| ImageNet map | 0.466 | 0.584 | 0.503–0.675 | 4.71 |
| Random orthogonal map | 0.020 | 0.316 | 0.190–0.463 | 3.25 |
| Orthogonal random directions | 0.000 | −0.005 | −0.014–0.008 | −0.57 |
| *Text vectors (T2)* | | | | |
| No mapping | 0.001 | 0.338 | 0.167–0.539 | 3.39 |
| Neutral-face map | 0.209 | 0.464 | 0.297–0.634 | 4.45 |
| ImageNet map | 0.287 | 0.533 | 0.374–0.682 | 4.82 |

**Activation-based mapping and five-emotion alignment provide complementary information.** We next explore whether combining activation-based mapping with emotion-based LOEO improves held-out correspondence. We first apply a fixed neutral-face or ImageNet mapping. Within each LOEO fold, we then fit an additional orthogonal transformation on five emotion pairs and evaluate the sixth. The minimum-change completion in Appendix B.7 keeps this additional transformation as close to the identity as the fit allows. Both vector sets use the same stimulus source, either AffectNet or T2. For each of 500 label permutations, we refit all six LOEO folds while keeping the initial activation-based mapping fixed.

Table 13 shows higher held-out cosine after combining the two fits. For AffectNet, additional LOEO raises cosine from 0.697 to 0.755 after neutral-face mapping and from 0.466 to 0.584 after ImageNet mapping, compared with 0.305 for LOEO alone. T2 shows the same pattern, increasing from 0.209 to 0.464 and from 0.287 to 0.533, compared with 0.338 for LOEO alone. Paired image activations therefore provide useful information that the five-emotion fit can build on. The combined procedure uses five emotion pairs and tests the excluded sixth. Only the initial activation-based map is fitted without emotion pairs.

The controls separate the role of the initial map from the relationships among the emotion vectors. A random orthogonal map followed by LOEO gives 0.316 on AffectNet, close to 0.305 for LOEO alone. This rotation preserves the source vectors' relationships. Replacing those vectors with independently sampled, matched-norm directions orthogonal to their corresponding targets instead gives near-zero LOEO cosine (−0.005). The combined fit improves held-out correspondence under this procedure. However, it does not change the source vectors' internal geometry, because orthogonal maps preserve within-source angles.

Table 14: Within-model steering for Qwen3-VL at image-token positions. diag and off are mean changes in log-probability for the injected emotion and the other five emotions, respectively. Random and orthogonal controls have matched norms.

| | $s = -0.5$ | | $s = -0.1$ | | $s = +0.1$ | | $s = +0.5$ | |
|---|---|---|---|---|---|---|---|---|
| Injected vector | diag | off | diag | off | diag | off | diag | off |
| *RAF-DB* | | | | | | | | |
| Image vector | $-4.94$ | $+1.32$ | $-1.60$ | $+0.42$ | $+1.85$ | $-0.36$ | $+5.34$ | $-1.55$ |
| Random | $+0.21$ | $+0.22$ | $+0.01$ | $+0.00$ | $+0.01$ | $+0.02$ | $+0.25$ | $+0.29$ |
| Orthogonal | $+0.27$ | $+0.20$ | $+0.02$ | $+0.00$ | $+0.00$ | $+0.02$ | $+0.20$ | $+0.31$ |
| *AffectNet* | | | | | | | | |
| Image vector | $-4.69$ | $+1.21$ | $-1.55$ | $+0.32$ | $+1.66$ | $-0.33$ | $+4.96$ | $-1.43$ |
| Random | $+0.19$ | $+0.16$ | $+0.01$ | $+0.00$ | $+0.01$ | $+0.02$ | $+0.19$ | $+0.24$ |
| Orthogonal | $+0.21$ | $+0.17$ | $+0.02$ | $+0.00$ | $+0.00$ | $+0.02$ | $+0.18$ | $+0.23$ |

**The need for mapping depends on the model pair.** Ministral-3 and Qwen3-VL provide a clear test of mapping because their native emotion vectors have nearly zero cosine. InternVL3.5 uses a Qwen3-based language backbone and already has higher native similarity with Qwen3-VL. Appendix E.3 includes this pair to test whether mapping also helps when the native vectors already correspond.

# E Causal Steering

Appendix D compares emotion vectors across sources, modalities, and architectures. We now explore whether their causal effects also transfer. We first test image-derived vectors within each model, then substitute text-derived vectors and vectors from other models. Further controls test effects on other readouts, generated continuations, injection positions, removal of the emotion-vector span, and intervention depth.

The main steering comparisons use a separate screening run to select held-out images that the target model classifies correctly, capped at $50$ per class across the six emotions and neutral. We use the same selected images when comparing interventions within a target model. This gives $274$ RAF-DB and $283$ AffectNet images for Qwen3-VL, $249$ and $243$ for InternVL3.5, and $253$ and $227$ for Ministral-3. Each intervention run recomputes its unsteered baseline, so measured baseline accuracy can differ from $100\%$. The controls below state any changes to the evaluation subset or readout.

## E.1 Within-Model Steering with Image-Derived Vectors

We first explore whether image-derived emotion vectors selectively change the same model's emotion predictions. We inject at all image-token positions. We report the mean change in the matched emotion's log-probability as diag and the mean change over the other five emotions as off. Neutral is a readout label but is excluded from both averages.

**Image emotion vectors selectively steer the readout in both directions.** To test whether the response follows intervention strength and sign, Table 14 reports the Qwen3-VL dose response. At $s = +0.5$, the matched emotion's log-probability increases by $5.34$ on RAF-DB and $4.96$ on AffectNet, while competing emotions decrease on average. Negative injection reverses this pattern, and the smaller interventions at $s = \pm 0.1$ give smaller effects with the same signs. This response supports a selective effect on the injected emotion.

We also inject matched-norm random and emotion-orthogonal directions to test whether an arbitrary perturbation produces the same response. Their mean effects have absolute values of at most $0.31$ and do not show the same separation between matched and competing emotions. The comparison therefore concerns selective steering, rather than requiring every control effect to be exactly zero. Figure 18 summarizes these effects alongside the text-vector results in Appendix E.2.

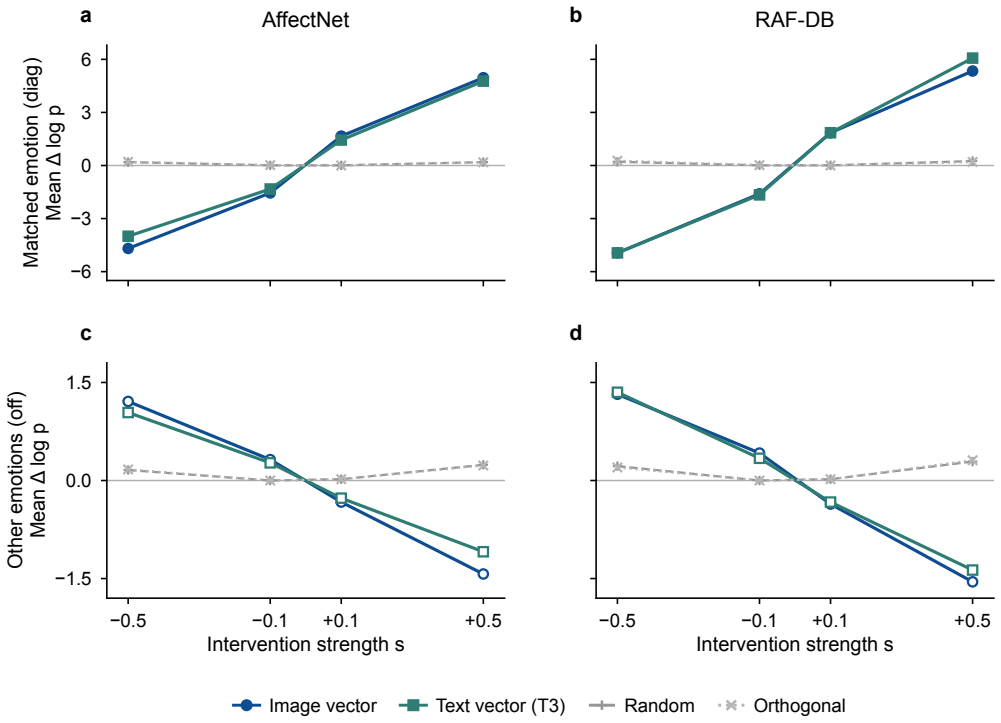



Figure 18: Image- and text-vector dose responses in Qwen3-VL. Columns compare AffectNet and RAF-DB. The upper row shows mean changes in the injected emotion's log-probability (diag), and the lower row shows the mean over the other five emotions (off). The two rows use different vertical scales. Points show the means in Tables 14 and 16; lines connect tested strengths. Image and T3 text vectors are injected at image-token positions. Random and orthogonal controls have matched norms.

Table 15: Within-model steering across the three core models. diag and off are mean changes in log-probability for the injected emotion and the other five emotions. Controls have matched norms.

| | RAF-DB | | | | AffectNet | | | |
|---|---|---|---|---|---|---|---|---|
| | $s = -0.5$ | | $s = +0.5$ | | $s = -0.5$ | | $s = +0.5$ | |
| Injected vector | diag | off | diag | off | diag | off | diag | off |
| *Qwen3-VL* | | | | | | | | |
| Image vector | −4.94 | +1.32 | +5.34 | −1.55 | −4.69 | +1.21 | +4.96 | −1.43 |
| Random | +0.21 | +0.22 | +0.25 | +0.29 | +0.19 | +0.16 | +0.19 | +0.24 |
| Orthogonal | +0.27 | +0.20 | +0.20 | +0.31 | +0.21 | +0.17 | +0.18 | +0.23 |
| *InternVL3.5* | | | | | | | | |
| Image vector | −1.93 | +0.51 | +1.88 | −0.40 | −2.07 | +0.45 | +2.19 | −0.52 |
| Random | +0.02 | +0.05 | +0.07 | +0.07 | +0.01 | +0.06 | +0.11 | +0.11 |
| Orthogonal | +0.04 | +0.05 | +0.05 | +0.07 | +0.03 | +0.06 | +0.09 | +0.11 |
| *Ministral-3* | | | | | | | | |
| Image vector | −0.88 | +0.24 | +1.25 | −0.21 | −0.97 | +0.22 | +1.29 | −0.19 |
| Random | −0.01 | +0.04 | +0.05 | −0.02 | −0.01 | +0.03 | +0.02 | −0.02 |
| Orthogonal | 0.00 | +0.04 | +0.04 | −0.02 | 0.00 | +0.03 | +0.02 | −0.02 |

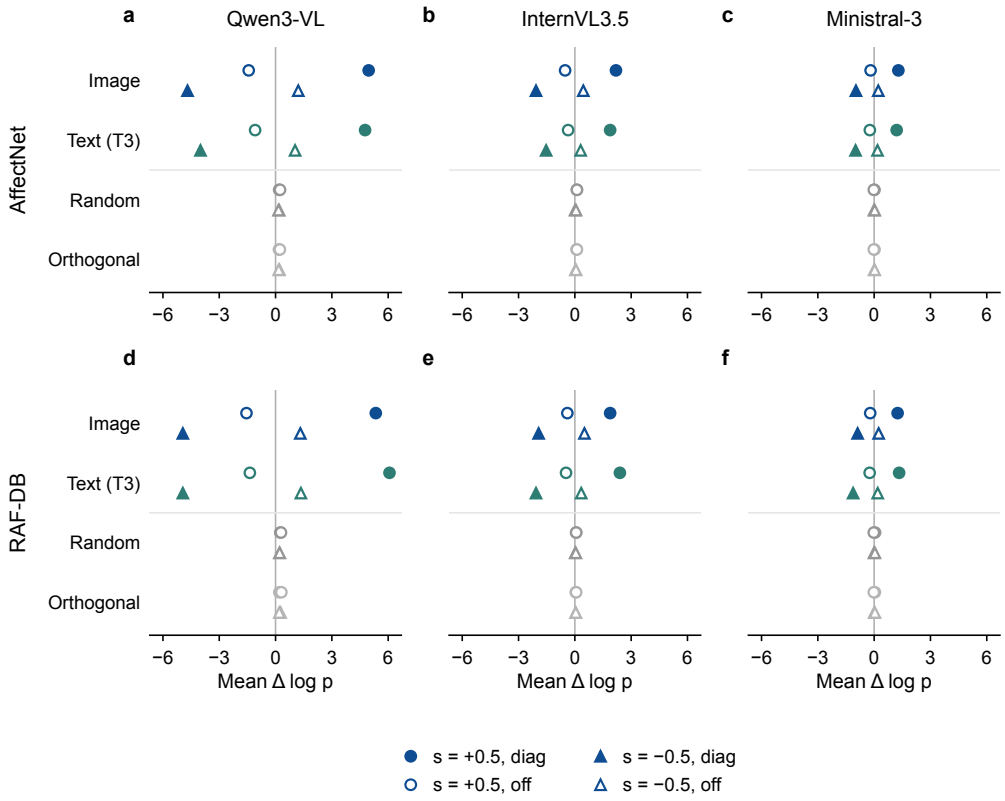



Figure 19: Within-model steering across three models and two image datasets. Image and T3 text vectors are compared with matched-norm random and orthogonal controls. The same horizontal scale is used throughout. Filled and open markers show diag and off; circles and triangles show $s = +0.5$ and $s = -0.5$. Points show the means in Tables 15 and 17.

**The selective effect replicates across architectures.** To test whether the pattern depends on Qwen3-VL, we repeat the intervention in InternVL3.5 and Ministral-3. Figure 19 shows the same selective sign reversal in both models and datasets, with exact values in Table 15. Absolute effects are smaller than in Qwen3-VL, but positive injection still raises the matched emotion and suppresses competing emotions on average. The random and orthogonal controls do not reproduce this pattern.

### E.2 Cross-Modal Steering with Text-Derived Vectors

**Text-derived emotion vectors retain causal function in the image stream.** To test transfer across modalities, we replace the image-derived vector with the same model's T3 text-derived vector and keep the rest of the protocol unchanged. No cross-modal alignment is applied. Qwen3-VL's T3 vectors have cosine similarity of only $0.24$ with RAF-DB vectors and $0.25$ with AffectNet vectors (Table 10). Nevertheless, Figure 18 shows selective steering in both directions, with exact values in Table 16. At $s = +0.5$, text-derived vectors raise the matched emotion's log-probability by $6.07$ on RAF-DB and $4.77$ on AffectNet, close to the corresponding image-vector effects.

**Text-derived vectors also steer images in the other two models.** We repeat the text-vector substitution in InternVL3.5 and Ministral-3. Figure 19 shows the same bidirectional selective pattern across both datasets, with effects comparable to those of each model's image-derived vectors (Table 17). Together with the LOEO results in Appendix D.2, this shows that low cosine similarity can coexist with both cross-modal emotion relationships and similar causal effects.

Table 16: Steering Qwen3-VL with its T3 text-derived vectors at image-token positions, without cross-modal alignment. diag and off are mean changes in log-probability for the injected emotion and the other five emotions.

| | $s = -0.5$ | | $s = -0.1$ | | $s = +0.1$ | | $s = +0.5$ | |
|---|---|---|---|---|---|---|---|---|
| Evaluation set | diag | off | diag | off | diag | off | diag | off |
| RAF-DB | $-4.94$ | $+1.35$ | $-1.66$ | $+0.34$ | $+1.85$ | $-0.33$ | $+6.07$ | $-1.37$ |
| AffectNet | $-4.00$ | $+1.04$ | $-1.33$ | $+0.27$ | $+1.44$ | $-0.27$ | $+4.77$ | $-1.09$ |

Table 17: Text- and image-vector steering within each model. T3 text vectors are injected at image-token positions without alignment. Image vectors provide the within-model reference. diag and off are mean changes in log-probability for the injected emotion and the other five emotions.

| | RAF-DB | | | | AffectNet | | | |
|---|---|---|---|---|---|---|---|---|
| | $s = -0.5$ | | $s = +0.5$ | | $s = -0.5$ | | $s = +0.5$ | |
| Injected vector | diag | off | diag | off | diag | off | diag | off |
| *Qwen3-VL* | | | | | | | | |
| Text vector (T3) | $-4.94$ | $+1.35$ | $+6.07$ | $-1.37$ | $-4.00$ | $+1.04$ | $+4.77$ | $-1.09$ |
| Image vector | $-4.94$ | $+1.32$ | $+5.34$ | $-1.55$ | $-4.69$ | $+1.21$ | $+4.96$ | $-1.43$ |
| *InternVL3.5* | | | | | | | | |
| Text vector (T3) | $-2.07$ | $+0.34$ | $+2.40$ | $-0.47$ | $-1.53$ | $+0.31$ | $+1.88$ | $-0.36$ |
| Image vector | $-1.93$ | $+0.51$ | $+1.88$ | $-0.40$ | $-2.07$ | $+0.45$ | $+2.19$ | $-0.52$ |
| *Ministral-3* | | | | | | | | |
| Text vector (T3) | $-1.12$ | $+0.19$ | $+1.33$ | $-0.24$ | $-0.99$ | $+0.18$ | $+1.20$ | $-0.23$ |
| Image vector | $-0.88$ | $+0.24$ | $+1.25$ | $-0.21$ | $-0.97$ | $+0.22$ | $+1.29$ | $-0.19$ |

### E.3 Cross-Architecture Steering with Mapped Vectors

We next explore whether mappings that recover vector correspondence also transfer causal effects across architectures. We use the activation-based mappings in Appendix D.3, learned without the six emotion vectors or their labels. Figure 20 compares direct injection with neutral-face and ImageNet mapping for all six ordered model pairs, with exact values in Table 18. We test both image-derived and T2 text-derived vectors.

**Neutral-face mapping restores selective transfer for Ministral-3 pairs.** The four ordered pairs involving Ministral-3 test transfer when native cosine is nearly zero. As shown in Table 18, direct injection gives little selective effect. Neutral-face mapping raises image-vector cosine to $0.64$–$0.72$ and restores bidirectional steering. For Ministral-3→Qwen3-VL at $s = +0.5$, the matched-emotion effect rises from $0.04$ without mapping to $4.11$, while the competing-emotion mean changes from $+0.18$ to $-0.77$. Negative injection reverses the selective response.

**ImageNet mapping also restores selective transfer.** To test whether transfer requires alignment images from an emotion dataset, we fit the maps to $3{,}000$ generic ImageNet images. Across the four Ministral-3 pairings, mapped image-vector cosine is lower than with neutral faces ($0.39$–$0.49$ versus $0.64$–$0.72$), but the selective sign reversal remains. Thus, paired images outside the emotion datasets also provide a useful map for causal transfer.

**Mappings learned from images also transfer text-derived vectors.** We further test transfer across both architectures and modalities by applying the image-based maps to text-derived vectors. Across the four Ministral-3 pairings, mapped text-vector cosine is only $0.18$–$0.29$, yet Table 18 shows selective steering under both injection signs. The causal effect therefore transfers even when correspondence between individual mapped and target vectors remains modest.

**Mapping does not improve every model pair.** Qwen3-VL and InternVL3.5 test whether mapping helps when native vectors already correspond. Before alignment, their image vectors have cosine

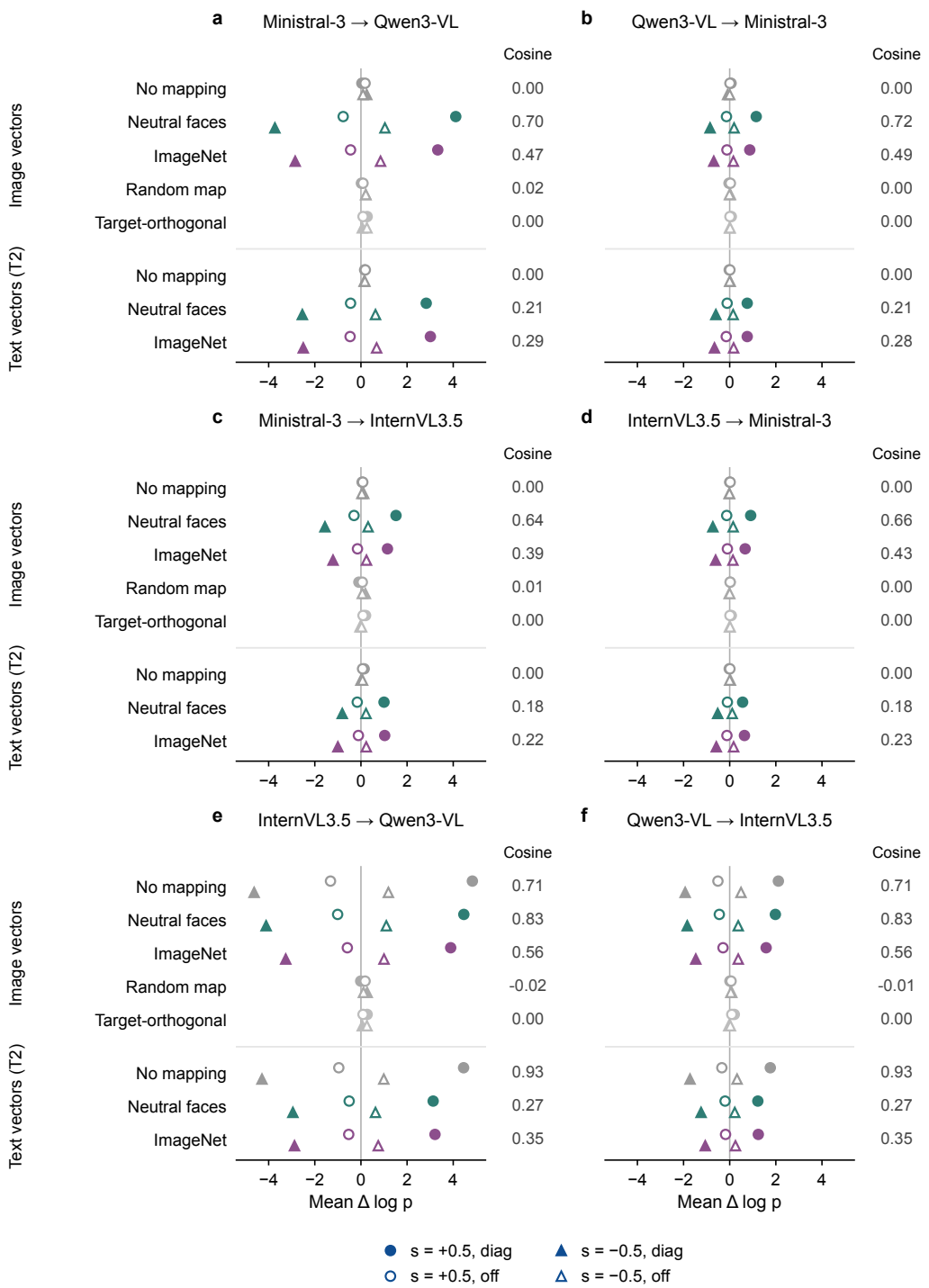



Figure 20: Cross-architecture steering on AffectNet for all six directed model pairs. Image and T2 text vectors are compared under no mapping, neutral-face mapping, and ImageNet mapping, with random orthogonal maps and target-orthogonal directions as image-vector controls. Filled/open markers show diag/off, and circles/triangles show positive/negative injection at $|s| = 0.5$. All comparisons use the same effect scale. The right-hand columns give cosine similarity to the target vectors of the same modality and source. Values are the means reported in Table 18.

Table 18: Cross-architecture steering on AffectNet. Source image or T2 text vectors are injected at the target's image-token positions. Cosine compares mapped vectors with target vectors of the same modality and source. diag and off are mean $\Delta \log p$ for the injected emotion and the other five emotions. Neutral-face and ImageNet maps use neither the six emotion vectors nor their labels.

| | | | $s = -0.5$ | | $s = +0.5$ | |
|---|---|---|---|---|---|---|
| Vector source | Mapping / control | Cosine | diag | off | diag | off |
| *Ministral-3 → Qwen3-VL* | | | | | | |
| Image | No mapping | 0.00 | +0.25 | +0.09 | +0.04 | +0.18 |
| | Random orthogonal map | 0.02 | +0.21 | +0.21 | +0.02 | +0.09 |
| | Target-orthogonal directions | 0.00 | +0.06 | +0.26 | +0.27 | +0.09 |
| | Neutral-face map | 0.70 | −3.73 | +1.04 | +4.11 | −0.77 |
| | ImageNet map | 0.47 | −2.85 | +0.85 | +3.33 | −0.45 |
| Text (T2) | No mapping | 0.00 | +0.14 | +0.17 | +0.16 | +0.19 |
| | Neutral-face map | 0.21 | −2.54 | +0.63 | +2.83 | −0.45 |
| | ImageNet map | 0.29 | −2.50 | +0.68 | +3.01 | −0.47 |
| *Qwen3-VL → Ministral-3* | | | | | | |
| Image | No mapping | 0.00 | −0.09 | +0.00 | +0.06 | +0.01 |
| | Random orthogonal map | 0.00 | +0.03 | −0.01 | −0.03 | +0.03 |
| | Target-orthogonal directions | 0.00 | −0.01 | +0.02 | +0.07 | +0.01 |
| | Neutral-face map | 0.72 | −0.85 | +0.19 | +1.15 | −0.14 |
| | ImageNet map | 0.49 | −0.69 | +0.16 | +0.87 | −0.12 |
| Text (T2) | No mapping | 0.00 | +0.02 | +0.00 | −0.02 | +0.01 |
| | Neutral-face map | 0.21 | −0.60 | +0.15 | +0.76 | −0.11 |
| | ImageNet map | 0.28 | −0.66 | +0.17 | +0.76 | −0.15 |
| *Ministral-3 → InternVL3.5* | | | | | | |
| Image | No mapping | 0.00 | +0.12 | +0.04 | +0.05 | +0.08 |
| | Random orthogonal map | 0.01 | +0.19 | +0.06 | −0.09 | +0.06 |
| | Target-orthogonal directions | 0.00 | −0.04 | +0.03 | +0.20 | +0.09 |
| | Neutral-face map | 0.64 | −1.56 | +0.31 | +1.52 | −0.30 |
| | ImageNet map | 0.39 | −1.21 | +0.24 | +1.15 | −0.15 |
| Text (T2) | No mapping | 0.00 | 0.00 | +0.07 | +0.14 | +0.07 |
| | Neutral-face map | 0.18 | −0.81 | +0.22 | +1.00 | −0.16 |
| | ImageNet map | 0.22 | −1.00 | +0.23 | +1.03 | −0.11 |
| *InternVL3.5 → Ministral-3* | | | | | | |
| Image | No mapping | 0.00 | −0.02 | −0.01 | +0.00 | +0.02 |
| | Random orthogonal map | 0.00 | −0.02 | 0.00 | +0.02 | +0.02 |
| | Target-orthogonal directions | 0.00 | −0.01 | +0.02 | +0.07 | +0.01 |
| | Neutral-face map | 0.66 | −0.73 | +0.15 | +0.91 | −0.13 |
| | ImageNet map | 0.43 | −0.61 | +0.14 | +0.67 | −0.10 |
| Text (T2) | No mapping | 0.00 | +0.03 | 0.00 | −0.02 | +0.01 |
| | Neutral-face map | 0.18 | −0.52 | +0.11 | +0.56 | −0.10 |
| | ImageNet map | 0.23 | −0.58 | +0.17 | +0.64 | −0.12 |
| *InternVL3.5 → Qwen3-VL* | | | | | | |
| Image | No mapping | 0.71 | −4.63 | +1.19 | +4.83 | −1.32 |
| | Random orthogonal map | −0.02 | +0.27 | +0.11 | −0.01 | +0.18 |
| | Target-orthogonal directions | 0.00 | +0.06 | +0.26 | +0.27 | +0.09 |
| | Neutral-face map | 0.83 | −4.11 | +1.10 | +4.46 | −1.01 |
| | ImageNet map | 0.56 | −3.26 | +1.00 | +3.89 | −0.59 |
| Text (T2) | No mapping | 0.93 | −4.29 | +0.99 | +4.45 | −0.96 |
| | Neutral-face map | 0.27 | −2.95 | +0.63 | +3.13 | −0.51 |
| | ImageNet map | 0.35 | −2.88 | +0.75 | +3.21 | −0.53 |
| *Qwen3-VL → InternVL3.5* | | | | | | |
| Image | No mapping | 0.71 | −1.93 | +0.49 | +2.11 | −0.51 |
| | Random orthogonal map | −0.01 | +0.09 | +0.05 | +0.01 | +0.06 |
| | Target-orthogonal directions | 0.00 | −0.04 | +0.03 | +0.20 | +0.09 |
| | Neutral-face map | 0.83 | −1.84 | +0.37 | +1.98 | −0.45 |
| | ImageNet map | 0.56 | −1.47 | +0.37 | +1.58 | −0.29 |
| Text (T2) | No mapping | 0.93 | −1.72 | +0.32 | +1.76 | −0.34 |
| | Neutral-face map | 0.27 | −1.24 | +0.22 | +1.22 | −0.20 |
| | ImageNet map | 0.35 | −1.06 | +0.25 | +1.24 | −0.18 |

Table 19: Emotion and gender readouts under Qwen3-VL steering at $s = +0.5$. Flip rate (%) compares predictions with the unsteered baseline. Accuracy is a fraction, and Mean averages the six injections.

| | Emotion readout | | Gender readout | |
|---|---|---|---|---|
| Injected emotion | Flip rate | Accuracy | Flip rate | Accuracy |
| No intervention | — | 0.982 | — | 0.961 |
| Anger | 79% | 0.204 | 2% | 0.969 |
| Disgust | 13% | 0.883 | 3% | 0.937 |
| Fear | 54% | 0.460 | 0% | 0.961 |
| Happiness | 38% | 0.628 | 4% | 0.925 |
| Sadness | 62% | 0.365 | 0% | 0.957 |
| Surprise | 38% | 0.620 | 1% | 0.949 |
| Mean | **47%** | **0.527** | **2%** | **0.950** |

similarity 0.71 and their T2 text vectors have cosine similarity 0.93. Direct injection already gives selective steering in both directions. Neutral-face mapping increases image-vector cosine but does not strengthen steering. Both maps reduce the cosine and steering strength of the text vectors. These results show that higher mapped cosine does not always imply a larger causal effect, and that mapping can be unnecessary for an already aligned pair.

Across the six ordered pairs, mapping is most useful when native vector similarity and direct transfer are weak. The recovered effects support shared causal function across models, while the differences between mapping conditions show that geometric similarity and steering strength remain distinct measures.

### E.4 Specificity, Coherence, and Injection Position

We next test whether steering selectively changes emotion-related predictions while preserving other model behavior. We compare emotion and gender readouts, check generated continuations for repetition, and vary the injection positions. These controls address different possible explanations for the observed change in emotion log-probabilities.

**Emotion steering largely preserves the gender readout.** To test effects on an unrelated visual attribute, we keep the $s = +0.5$ image-token intervention and change the readout question from emotion to gender. Both tasks use the RAF-DB images. We measure prediction flips relative to each task's unsteered baseline and accuracy against its dataset labels.

Table 19 shows a much larger change in emotion predictions than in gender predictions. Averaged over injections, emotion predictions change on 47% of images and accuracy falls from 0.982 to 0.527. Gender predictions change on only 2%, with accuracy remaining close to baseline (0.961 to 0.950). This supports selectivity relative to the gender readout, although the two reported averages use slightly different evaluation subsets.

**Decoded continuations remain coherent and follow the induced interpretation.** To test whether steering changes more than the first emotion word, we decode full continuations using Qwen3-VL's image-derived vectors. Across 630 continuations per dataset, distinct-1, the fraction of unique generated tokens, averages 0.92 on RAF-DB and 0.93 on AffectNet and never falls below 0.73. The longest repeated-token run is one token even at $s = \pm 0.5$. These diagnostics show no repetition collapse in the tested outputs.

Figure 21 shows selected continuations that name the injected emotion and explain the answer using visual cues. However, the disgust example is a rare success. Across 3,500 tested AffectNet validation images, disgust steering changes the readout to the target on only 32 images, under 1%. The other five emotions succeed on 14–60% of the images tested. Thus, a mean log-probability shift does not guarantee frequent changes in generated emotion labels. Disgust also has low classification recall in Table 2, although classification and induced generation measure different abilities.

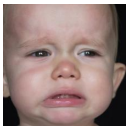
| Image | Continuation, unsteered above and steered below |
|---|---|
|  | **none**: extreme **fear**, **shock**, or **terror**. Here's what in the image shows it: *Wide, bulging eyes*: The eyes are opened very wide, with... **anger**: extreme **anger** or **rage**. Here's what in the image shows it: Eyes: They are wide open and staring intensely, which can indicate **shock**, **fear**, or extreme **anger**. Eyebrows: The *eyebrows are drawn together and furrowed*, creating deep lines on the... |
|  | **none**: **sad** or **upset**, and the image shows this through their *tear-streaked face, downturned mouth, and furrowed brow*. **fear**: **scared** or **frightened**. The image shows this through several classic signs of **fear** in a child: - *Wide, dilated pupils* (visible in the eyes) — this is a physiological response to **fear** or **surprise**, as the body prepares to react. - *Tense facial muscles*... |
|  | **none**: **uncomfortable**, **irritated**, or possibly **sick**. Here's what in the image shows this: 1. The Action: The child is actively... **happiness**: **happy** or **content**. Here's what in the image shows it: Smiling Mouth: The person's *lips are curved upwards in a clear, gentle smile*. Glowing Eyes: Their *eyes are bright and crinkled at the corners*, which is often associated with genuine... |
|  | **none**: **annoyed** or **irritated**. Here's what in the image shows it: - Eyebrows: Their *eyebrows are drawn together and slightly raised*, creating... **sadness**: **sad** or **disappointed**. The image shows this through their facial expression: their *eyebrows are slightly furrowed*, their *eyes are downcast*, and their *mouth is turned downward in a frown*. These are all classic signs of **sadness** or **disappointment**. |
|  | **none**: **intense** and **focused**, with a hint of **vulnerability** or **intensity**. Here's what in the image shows it: - Eyes: Their *eyes are wide open*... **surprise**: **surprised** or **startled**. Here's what in the image shows it: - Wide, Alert Eyes: The most telling feature is the person's eyes. *They are wide open, and the pupils are dilated*, which is a classic physiological response to being **startled** or **surprised**. The gaze... |
|  | **none**: **content** and **satisfied**. The image shows this through the person's *relaxed facial expression* — their *eyes are gently closed*, and their... **disgust**: **disgusted** or **repulsed**. Here's what in the image shows it: Facial Expression: The person's *eyes are squeezed shut*, and their *nose is wrinkled*. This is a classic, involuntary reaction to something unpleasant or foul-tasting. Action: They are... |

Figure 21: Selected Qwen3-VL continuations under its own AffectNet emotion vectors at image-token positions and $s = +0.5$. Each block pairs a different validation image with one injected emotion and shows unsteered text above steered text. Green and red highlights mark unsteered and steered text, respectively. Bold marks labels and emotion terms, and italics mark selected visual descriptions. Ellipses mark truncation.

Table 20: Injection-position comparison for Qwen3-VL at $s = +0.5$, with vectors held fixed. diag and off are mean changes in log-probability for the injected emotion and the other five emotions.

| | RAF-DB | | AffectNet | |
|---|---|---|---|---|
| Injection positions | diag | off | diag | off |
| Image tokens (default) | $+5.34$ | $-1.55$ | $+4.96$ | $-1.43$ |
| Question tokens | $+2.50$ | $+0.70$ | $+2.27$ | $+0.13$ |
| Answer prefix | $+4.71$ | $-2.31$ | $+4.73$ | $-1.95$ |
| All positions | $+5.23$ | $-5.51$ | $+5.55$ | $-5.57$ |

Table 21: Emotion recognition accuracy after partial ($f = 0.5$) or full ($f = 1$) removal in Qwen3-VL at image-token positions. Classification includes the six emotions and neutral. Each random control matches the dimension of the emotion direction or span directly above it.

| | RAF-DB | | AffectNet | |
|---|---|---|---|---|
| Removed direction or span | $f = 0.5$ | $f = 1$ | $f = 0.5$ | $f = 1$ |
| No removal ($f = 0$) | 0.993 | | 0.986 | |
| Single emotion vector (rank 1) | 0.969 | 0.946 | 0.950 | 0.902 |
| Random direction | 0.991 | 0.991 | 0.988 | 0.987 |
| Centered emotion-vector span (rank 5) | 0.938 | **0.814** | 0.901 | **0.781** |
| Random subspace | 0.990 | 0.988 | 0.988 | 0.990 |

**Steering strength and selectivity depend on injection position.** To explore how the intervention location affects steering, we keep the vector and strength fixed and vary the token positions. Table 20 shows that question-token injection gives a weaker matched-emotion effect and does not suppress competing emotions on average. Answer-prefix injection gives a matched effect close to image-token injection, while all-position injection suppresses competing emotions most strongly. The vectors therefore remain effective at several locations, with different strengths and effects on competing labels.

We use image-token injection as the default because it produces selective steering and places the intervention at the visual-input positions. This specifies where we change the residual stream, without implying that later effects remain limited to those positions.

### E.5 Removing the Emotion-Vector Span

The steering experiments show that adding emotion vectors can change emotion predictions. We next explore whether removing their components from hidden states impairs normal emotion recognition. We compare individual emotion directions with the centered span of the six source-specific vectors. We obtain an orthonormal basis for the span by re-centering the vectors and applying SVD. Each condition is compared with a random control of the same dimension.

Let $\mathbf{B}$ contain an orthonormal basis for the selected direction or subspace. At the same layer and image-token positions as in the steering experiments, we remove a fraction $f$ of each hidden state's projection onto it,

$$\boldsymbol{h} \leftarrow \boldsymbol{h} - f\,\boldsymbol{h}\mathbf{B}\mathbf{B}^{\top}, \qquad f \in [0, 1]. \tag{16}$$

Here, $f = 0.5$ removes half of the projected component, and $f = 1$ removes it completely. For single-vector removal, we remove each of the six emotion directions separately and average the resulting accuracies.

**Removing the emotion-vector span reduces recognition accuracy.** As shown in Table 21, full removal of the centered emotion-vector span lowers accuracy from $0.993$ to $0.814$ on RAF-DB and from $0.986$ to $0.781$ on AffectNet. Partial removal and removal of a single emotion direction cause smaller decreases. In contrast, the random controls leave accuracy close to baseline under both removal strengths. This comparison supports a contribution of the emotion-vector span beyond the effect of removing arbitrary directions. However, removing the span does not eliminate emotion

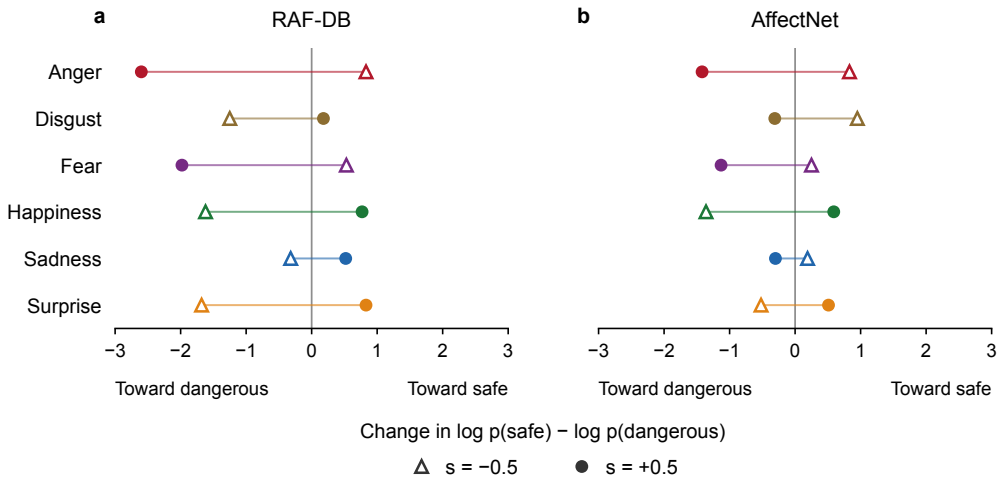



Figure 22: Safe–dangerous readout under Qwen3-VL emotion steering. Points show mean changes from the unsteered model, with paired markers for $s = -0.5$ (open triangles) and $s = +0.5$ (filled circles). Negative values shift the prediction toward dangerous and positive values toward safe.

Table 22: Change in the safe–dangerous log-probability contrast under Qwen3-VL emotion steering at image-token positions. Positive values favor *safe*, and negative values favor *dangerous*.

| | RAF-DB | | AffectNet | |
|---|---|---|---|---|
| Injected emotion | $s = -0.5$ | $s = +0.5$ | $s = -0.5$ | $s = +0.5$ |
| Anger | $+0.83$ | $-2.60$ | $+0.83$ | $-1.42$ |
| Disgust | $-1.25$ | $+0.18$ | $+0.95$ | $-0.31$ |
| Fear | $+0.53$ | $-1.98$ | $+0.25$ | $-1.13$ |
| Happiness | $-1.62$ | $+0.77$ | $-1.36$ | $+0.59$ |
| Sadness | $-0.32$ | $+0.52$ | $+0.19$ | $-0.30$ |
| Surprise | $-1.68$ | $+0.83$ | $-0.52$ | $+0.51$ |

recognition. The extracted directions therefore contribute to normal recognition but do not capture all of the information that supports it.

### E.6 Beyond the Emotion Readout

We further explore whether emotion steering affects a related prediction that uses no emotion label. We also test whether the induced emotion is associated specifically with the person in the image by changing who the question refers to.

**Emotion steering also changes the safe–dangerous prediction.** We replace the emotion question with a binary *safe–dangerous* prediction while keeping the image and intervention unchanged. Figure 22 shows the change in $\log p(\text{safe}) - \log p(\text{dangerous})$ relative to the unsteered model, with exact values in Table 22.

Positive anger and fear injections move the prediction toward *dangerous* on both datasets, while happiness moves it toward *safe*. Negative injection reverses these effects. Disgust and sadness show less consistent changes across datasets. Together with the largely unchanged gender readout, this suggests that steering can affect a related prediction beyond the explicit emotion label.

**Steering is strongest for the depicted person but also affects other subjects.** To test who the induced emotion is attributed to, we keep the image and intervention fixed and ask about the depicted person, the assistant, or a third party absent from the image.

Table 23: Effect of the queried subject on Qwen3-VL emotion steering at image-token positions. diag and off are mean changes in log-probability for the injected emotion and the other five emotions.

| | RAF-DB | | | | AffectNet | | | |
|---|---|---|---|---|---|---|---|---|
| | $s = -0.5$ | | $s = +0.5$ | | $s = -0.5$ | | $s = +0.5$ | |
| Emotion attributed to | diag | off | diag | off | diag | off | diag | off |
| the depicted person | $-3.43$ | $+0.30$ | $\mathbf{+2.63}$ | $-1.10$ | $-2.83$ | $+0.34$ | $\mathbf{+2.37}$ | $-0.83$ |
| the assistant | $-1.94$ | $+0.37$ | $+1.78$ | $-0.53$ | $-1.73$ | $+0.26$ | $+1.61$ | $-0.44$ |
| a third party | $-2.20$ | $+0.31$ | $+1.91$ | $-0.60$ | $-1.68$ | $+0.25$ | $+1.63$ | $-0.38$ |

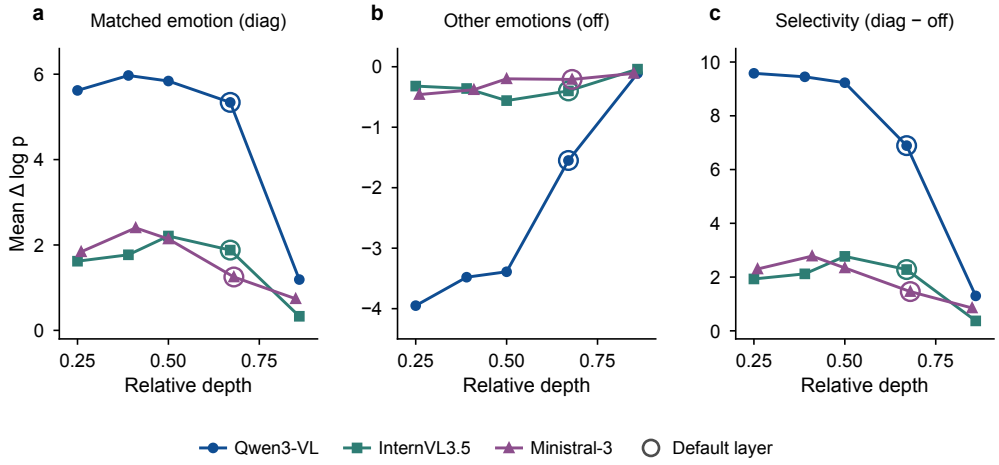



Figure 23: Depth dependence of steering on RAF-DB at $s = +0.5$. Mean $\Delta \log p$ is shown for the matched emotion (diag), the other five emotions (off), and their difference. Lines connect measured layers. Rings mark the default layers. All depths use the same residual-norm scaling rule. Table 24 gives exact values.

Table 23 shows the strongest matched-emotion effects when the question refers to the depicted person. However, steering also changes answers about the assistant and the third party. The effect is therefore partly tied to the visual subject, but can also influence emotion predictions about other subjects in the prompt.

### E.7 Causal Effect Across Depth

We choose the default intervention layer at approximately the same relative depth as the representational analyses. To explore whether this choice is also where steering is strongest, we repeat the intervention at approximately matched fractional depths in all three models (Figure 23; Table 24).

**Steering is stronger before the default layer and weaker near the output.** Figure 23 shows a broad region of strong effects at relative depths 0.25–0.50 in Qwen3-VL. InternVL3.5 peaks at 0.50 and Ministral-3 at 0.41, and all three have weaker effects at the deepest tested layer. At the default layer, selectivity is about 28%, 18%, and 47% below the respective model maxima. The reported effects therefore do not rely on selecting the strongest intervention layer. Because all depths use residual-norm scaling, these results describe steering under that calibration and do not give a scale-independent ranking of layers.

Table 24: Steering across depth on RAF-DB at $s = +0.5$. diag and off are mean $\Delta \log p$ for the injected emotion and the other five emotions. Selectivity is diag minus off. Bold marks each model's largest diag and selectivity. Differences use unrounded values.

| Layer | Relative depth | diag | off | Selectivity (diag − off) |
|---|---|---|---|---|
| *Qwen3-VL* | | | | |
| 9 | 0.25 | +5.62 | −3.95 | **+9.58** |
| 14 | 0.39 | **+5.97** | −3.48 | +9.45 |
| 18 | 0.50 | +5.84 | −3.39 | +9.23 |
| 24 | 0.67 | +5.34 | −1.55 | +6.89 |
| 31 | 0.86 | +1.19 | −0.11 | +1.30 |
| *InternVL3.5* | | | | |
| 9 | 0.25 | +1.62 | −0.32 | +1.93 |
| 14 | 0.39 | +1.77 | −0.36 | +2.12 |
| 18 | 0.50 | **+2.21** | −0.56 | **+2.77** |
| 24 | 0.67 | +1.88 | −0.40 | +2.28 |
| 31 | 0.86 | +0.33 | −0.04 | +0.37 |
| *Ministral-3* | | | | |
| 9 | 0.26 | +1.84 | −0.46 | +2.30 |
| 14 | 0.41 | **+2.40** | −0.38 | **+2.78** |
| 17 | 0.50 | +2.14 | −0.20 | +2.34 |
| 23 | 0.68 | +1.25 | −0.21 | +1.46 |
| 29 | 0.85 | +0.74 | −0.11 | +0.85 |

# F Shared Emotion Subspace

The preceding analyses show that emotion vectors can retain similar relationships and causal effects despite differences in source and model coordinates. We now explore whether averaging aligned vectors across models and sources yields a shared emotion subspace. We first construct this subspace, then test how much of the original vectors and their causal effects it retains. Lexical readouts, generated continuations, and held-out Gemma-3 experiments provide further checks.

## F.1 Separating Shared Emotion Structure from Source and Model Variation

Let $\boldsymbol{v}_e^{m,q}$ denote the emotion vector for model $m$, stimulus source $q$, and emotion $e \in \mathcal{E}$, written as a row vector. The matrix $\mathbf{V}_{m,q} \in \mathbb{R}^{|\mathcal{E}| \times D}$ stacks these vectors as rows in the same emotion order for every model and source, where $D$ is the residual width.

We extend the pairwise alignment in Equation 3 to the three core models using generalized Procrustes alignment. Let $\widetilde{\mathbf{X}}_m$ contain centered activations from model $m$, with rows paired across models using the same neutral-face or ImageNet images. For the set of three core models $\mathcal{M}$, we jointly fit orthogonal transformations $\mathbf{R}_m$ from each model to a common coordinate frame and a shared activation template $\mathbf{G}$ by minimizing

$$\min_{\mathbf{G}, \{\mathbf{R}_m\}_{m \in \mathcal{M}}} \sum_{m \in \mathcal{M}} \left\| \widetilde{\mathbf{X}}_m \mathbf{R}_m - \mathbf{G} \right\|_F^2, \qquad \mathbf{R}_m^\top \mathbf{R}_m = \mathbf{I}. \tag{17}$$

We alternate between fitting each transformation to the current template and updating the template to the mean of the aligned activations. This fit uses neither the six emotion vectors nor their labels. We then apply each transformation to the emotion vectors,

$$\widetilde{\mathbf{V}}_{m,q} = \mathbf{V}_{m,q} \mathbf{R}_m, \qquad \widetilde{\boldsymbol{v}}_e^{m,q} = \boldsymbol{v}_e^{m,q} \mathbf{R}_m, \quad e \in \mathcal{E}. \tag{18}$$

The consensus matrix $\bar{\mathbf{V}} = \mathbb{E}_{m,q}[\widetilde{\mathbf{V}}_{m,q}]$ is the mean over models and sources. Its row for emotion $e$ is the consensus vector $\bar{\boldsymbol{v}}_e$ from Section 6. We write

$$\widetilde{\mathbf{V}}_{m,q} = \bar{\mathbf{V}} + \boldsymbol{\Delta}_q + \boldsymbol{\Delta}_m + \mathbf{E}_{m,q}. \tag{19}$$

The source term $\mathbf{\Delta}_q$ describes each source's difference from this consensus, averaged over models. The model term $\mathbf{\Delta}_m$ describes each model's difference, averaged over sources. The remaining term $\mathbf{E}_{m,q}$ contains source–model interactions and estimation residuals. We define the source and model terms as

$$\mathbf{\Delta}_q = \mathbb{E}_m\left[\widetilde{\mathbf{V}}_{m,q}\right] - \bar{\mathbf{V}}, \qquad \mathbf{\Delta}_m = \mathbb{E}_q\left[\widetilde{\mathbf{V}}_{m,q}\right] - \bar{\mathbf{V}}. \tag{20}$$

These definitions give zero mean across sources for $\mathbf{\Delta}_q$ and across models for $\mathbf{\Delta}_m$. The remaining term then has zero mean over either models or sources, separating the average source and model differences from their interactions.

Source and modality are related because each source contains either text or images. The variance analysis below keeps all model–source–emotion combinations and gives each source equal weight. Total variance is measured before averaging over sources. We obtain the emotion, model, and modality components by averaging over the other factors, and assign the remaining variance to residuals and interactions. This last term includes differences among sources within a modality. These scalar variance components summarize different aspects of the data from the six-emotion source deviations in Equation 19.

**Centered consensus vectors define the shared emotion subspace.** For each emotion $e \in \mathcal{E}$, we average its aligned vector over models and stimulus sources,

$$\bar{\boldsymbol{v}}_e = \mathbb{E}_{m,q}\left[\widetilde{\boldsymbol{v}}_e^{m,q}\right], \qquad \bar{\boldsymbol{v}} = \mathbb{E}_{e\in\mathcal{E}}\left[\bar{\boldsymbol{v}}_e\right]. \tag{21}$$

We center the rows of $\bar{\mathbf{V}}$ by subtracting their mean $\bar{\boldsymbol{v}}$ and define the shared emotion subspace as

$$\mathcal{S}_{\text{emo}} = \operatorname{span}\left\{\bar{\boldsymbol{v}}_e - \bar{\boldsymbol{v}}\right\}_{e\in\mathcal{E}}. \tag{22}$$

This subspace describes the shared differences among emotions. The six centered vectors sum to zero, which gives $\dim(\mathcal{S}_{\text{emo}}) \le 5$. This bound follows from the construction and is not a chosen number of principal components.

When an explicit basis is required for projection, we stack the centered consensus vectors as columns and compute

$$\mathbf{W}_{\text{emo}} = \operatorname{orth}\left(\left[(\bar{\boldsymbol{v}}_e - \bar{\boldsymbol{v}})^\top\right]_{e\in\mathcal{E}}\right). \tag{23}$$

The columns of $\mathbf{W}_{\text{emo}}$ are orthonormal and span $\mathcal{S}_{\text{emo}}$, so

$$\operatorname{col}(\mathbf{W}_{\text{emo}}) = \mathcal{S}_{\text{emo}}. \tag{24}$$

The corresponding orthogonal projector is

$$\mathbf{P}_{\text{emo}} = \mathbf{W}_{\text{emo}}\mathbf{W}_{\text{emo}}^\top. \tag{25}$$

Averaging over models and sources follows the marginalization principle of demixed PCA (Kobak et al., 2016). Here, however, we retain the full span of the centered consensus vectors rather than selecting a number of principal components. We treat this span as a candidate shared emotion subspace and test its geometric, causal, and lexical properties below.

**Shared differences among emotions coexist with source variation.** To measure the variation shared across conditions, Table 25 separates emotion-related variance from model, modality, and remaining variation. Emotion accounts for $0.42$–$0.47$ of variance within each model. After neutral-face alignment across models, it accounts for $0.326$, compared with $0.004$ for model and $0.003$ for modality. The residual and interaction terms account for $0.668$. Thus, the small variance shares for model and modality describe small average shifts across all emotions. They do not imply that source-specific or emotion-specific differences disappear.

We next measure how much of each aligned source-specific vector lies in the shared span by computing its cosine with its orthogonal projection onto $\mathcal{S}_{\text{emo}}$. Averaged over all seven sources and six emotions, the cosines are $0.64$, $0.59$, and $0.61$ for Qwen3-VL, Ministral-3, and InternVL3.5 under neutral-face alignment. The AffectNet-only values are $0.77$, $0.68$, and $0.76$, respectively. This measures reconstruction of the original vectors by the shared span, not similarity to the consensus vector for each emotion.

Table 25: Variance components of emotion vectors across all seven sources, weighted equally. Entries are fractions of total variance before source averaging. Cross-model estimates follow generalized Procrustes alignment. Residual/interactions includes within-modality source variation and interactions among factors. Values may not sum to one after rounding.

| Model / alignment | Emotion | Modality | Model | Residual / interactions |
|---|---|---|---|---|
| *Within each model* | | | | |
| Qwen3-VL | 0.474 | 0.004 | — | 0.522 |
| InternVL3.5 | 0.447 | 0.004 | — | 0.550 |
| Ministral-3 | 0.423 | 0.006 | — | 0.571 |
| *Across all three aligned models* | | | | |
| Neutral-face alignment | 0.326 | 0.003 | 0.004 | 0.668 |
| ImageNet alignment | 0.308 | 0.002 | 0.004 | 0.686 |

We also test whether the shared structure depends on using neutral faces for alignment. ImageNet alignment gives a similar variance pattern, with emotion accounting for $0.308$, model for $0.004$, and residuals and interactions for $0.686$. Mean projection cosine becomes $0.63$, $0.57$, and $0.59$ across all sources, and $0.69$, $0.58$, and $0.66$ on AffectNet, in the same model order. Generic images therefore support the shared subspace, although they give weaker reconstruction, especially for AffectNet vectors.

The consensus vectors also retain agreement with human affective ratings (Figure 3e; Appendix C.4). Against R&M, PC1–valence $|r|$ is $0.93$ under both alignment conditions. PC2–arousal $|r|$ is $0.96$ with neutral-face alignment and $0.97$ with ImageNet alignment. These results connect the shared emotion relationships to the human affective reference.

### F.2 Causal Localization of the Shared Emotion Subspace

We next explore whether the shared subspace retains the causal effect of Qwen3-VL's original emotion vectors. We first express its projector in Qwen3-VL's coordinates as $\mathbf{P}^{\mathrm{Q}}_{\mathrm{emo}} = \mathbf{R}_{\mathrm{Q}}\mathbf{P}_{\mathrm{emo}}\mathbf{R}_{\mathrm{Q}}^{\top}$. We then split each original vector $\boldsymbol{v}$ into its shared projection and orthogonal remainder,

$$\boldsymbol{v}_{\mathrm{shared}} = \boldsymbol{v}\mathbf{P}^{\mathrm{Q}}_{\mathrm{emo}}, \qquad \boldsymbol{v}_{\perp} = \boldsymbol{v}(\mathbf{I} - \mathbf{P}^{\mathrm{Q}}_{\mathrm{emo}}). \tag{26}$$

The shared projection is the part of the original vector inside the mapped shared subspace. The remainder is the part outside it. Neither is the consensus vector itself, which averages aligned vectors across models and sources. The remainder may contain source-dependent and emotion-related information, so we test its effect separately rather than treating it as noise.

**The shared projection retains the selective causal effect under both norm conditions.** To compare the causal effects of the directions themselves, we first normalize each component to unit length. As shown in Figure 24a and Table 26, the shared projection gives $(\mathrm{diag}, \mathrm{off}) = (+5.39, -1.62)$, close to or larger than the original vector's $(+4.96, -1.43)$. The remainder gives a smaller matched effect of $+1.00$ and weaker competing-emotion suppression of $-0.14$. The $95\%$ image-bootstrap intervals for diag are $[5.18, 5.61]$, $[4.77, 5.16]$, and $[0.94, 1.05]$ for the projection, original vector, and remainder. The projection's selectivity exceeds the original vector's by $0.62$, with a paired $95\%$ interval of $[0.59, 0.66]$.

Unit normalization changes the lengths of the original components. We therefore repeat the intervention without renormalizing either component. The projection still gives $(\mathrm{diag}, \mathrm{off}) = (+5.14, -1.47)$, while the remainder gives $(+0.68, -0.14)$. The shared projection therefore retains the selective effect under both norm conditions. Because the model response is nonlinear, the effects of separately injecting the two components need not sum to the original-vector effect.

To test whether the smaller remainder effect depends on emotion identity, we compare it with a random direction in the same orthogonal complement and with remainders assigned rotated emotion labels. At equal norm, these controls give $(\mathrm{diag}, \mathrm{off}) = (+0.19, +0.22)$ and $(+0.12, +0.03)$, respectively (Figure 24b). Neither reproduces the selective response of the real remainder. Emotion-

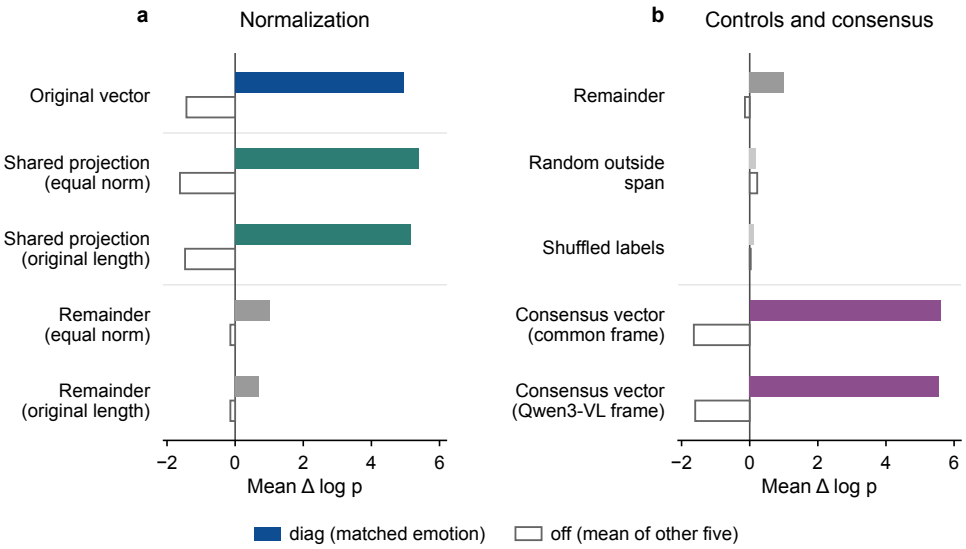



Figure 24: Shared-subspace steering in Qwen3-VL on AffectNet at $s = +0.5$. (a) Shared projection and remainder under both norm conditions, compared with the original vector. (b) Equal-norm remainder controls and consensus vectors under two reference frames. The equal-norm remainder is repeated for comparison. Bars show the means in Table 26; bootstrap intervals for selected comparisons are reported in the text.

Table 26: Shared-subspace steering in Qwen3-VL on AffectNet images at $s = +0.5$. The span uses all three core models and seven sources. Equal-norm directions have unit length. The next block retains the original component lengths. Shuffled labels rotate remainder labels. diag and off are mean $\Delta \log p$ for the injected emotion and the other five emotions.

| | Mean $\Delta \log p$ | |
|---|---|---|
| Injected vector | diag | off |
| *Equal-norm directions* | | |
| Original vector $\boldsymbol{v}$ | $+4.96$ | $-1.43$ |
| Shared projection $\boldsymbol{v}_{\text{shared}}$ | $+5.39$ | $-1.62$ |
| Remainder $\boldsymbol{v}_{\perp}$ | $+1.00$ | $-0.14$ |
| Random outside span $\mathcal{S}_{\text{emo}}$ | $+0.19$ | $+0.22$ |
| Shuffled labels (remainder) | $+0.12$ | $+0.03$ |
| Consensus vector (mapped to Qwen3-VL) | $+5.60$ | $-1.64$ |
| *Original component lengths* | | |
| Shared projection $\boldsymbol{v}_{\text{shared}}$ | $+5.14$ | $-1.47$ |
| Remainder $\boldsymbol{v}_{\perp}$ | $+0.68$ | $-0.14$ |
| *Reference-frame control* | | |
| Consensus vector (Qwen3-VL reference frame) | $+5.56$ | $-1.60$ |

related information therefore remains outside the shared span, although its effect is smaller than that of the projection.

**Consensus vectors themselves also steer Qwen3-VL.** We further test the consensus vectors themselves, rather than projections of Qwen3-VL's original vectors. After mapping back to Qwen3-VL, they give $(\text{diag}, \text{off}) = (+5.60, -1.64)$ in Table 26. Constructing the consensus with Qwen3-VL as the reference gives a similar result of $(+5.56, -1.60)$, as shown in Figure 24b. Thus, both constructions produce consensus vectors that retain selective steering.

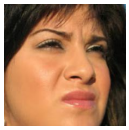
| Image | Injected | Continuation |
|---|---|---|
|  | **none** | **annoyed**, **skeptical**, or slightly **disgusted**. Here's what in the image shows it: 1. Eyebrows: Their *eyebrows are slightly raised* and drawn... |
| | **anger** | **angry** or **frustrated**. Here's what in the image shows it: - Eyebrows: Their *eyebrows are drawn together and slightly lowered*, a classic sign of **anger** or... |
| | **disgust** | **disgusted** or **repulsed**. Here's what in the image shows it: - Eyebrow and Eye Expression: Their *eyebrows are slightly furrowed*, and their *eyes are rolled upward* and... |
| | **fear** | **scared** or **frightened**. Here's what in the image shows it: - *Wide, darting eyes*: Her *eyes are wide open and looking sharply to the side* (her right), as if she's... |
| | **happiness** | **amused** or **amusedly skeptical**, likely at something funny or absurd. Here's what in the image shows it: 1. The Eyes: The most telling feature is the person's eyes... |
| | **sadness** | **sad** or **disappointed**. Here's what in the image shows it: - *Downward gaze and drooping eyelids*: The person's eyes are looking down and slightly to the side, and... |
| | **surprise** | **surprised** or **startled**. Here's what in the image shows it: - Eyes: Their *eyes are wide open and looking upward and to the side* (toward the upper right). This is a... |
| | **none** | **annoyed** or **frustrated**. Here's what in the image shows it: - Eyebrows: Her *eyebrows are drawn together and slightly furrowed*, a classic sign of... |
| | **anger** | **angry** or **frustrated**. Here's what in the image shows it: - Eyebrows: Their *eyebrows are drawn together and slightly furrowed*, a classic sign of **anger** or... |
| | **disgust** | **disgusted** or **repulsed**. Here's what in the image shows it: - Eyebrows: Their *eyebrows are drawn together and slightly raised*, forming a *furrowed brow*. This is a... |
|  | **fear** | **scared** or **frightened**. Here's what in the image shows it: - *Wide, dilated pupils*: The person's *eyes are wide open*, and their *pupils appear very large*. This is a... |
| | **happiness** | **happy** or **amused**. Here's what in the image shows it: - Smiling mouth: Their *lips are curved upwards in a clear, gentle smile*. - *Crinkled eyes*: The corners... |
| | **sadness** | **sad** or **upset**. Here's what in the image shows it: - Eyebrows: Their *eyebrows are drawn together and slightly furrowed*, a classic sign of **sadness**, **worry**, or... |
| | **surprise** | **surprised** or **shocked**. Here's what in the image shows it: - Wide, **startled** eyes: Her *eyes are wide open*, and her *pupils appear dilated*, which is a common... |

Figure 25: Selected continuations for two AffectNet validation images under all six consensus-vector injections in Qwen3-VL at $s = +0.5$. Both images have ground-truth label *disgust*. The first row for each image is unsteered. Highlight colors follow the injected emotions in Figure 3, while unsteered highlights remain black. Bold marks labels and emotion terms, and italics mark selected visual descriptions. The leading *They feel* is omitted, and ellipses mark truncation.

### F.3 Generated Continuations and Lexical Readouts of Consensus Vectors

**Selected continuations follow all six consensus-vector injections.** To explore whether consensus vectors change generated explanations as well as emotion scores, we apply each of the six injections to the same image. Figure 25 shows two images under consensus vectors mapped into Qwen3-VL. For each image, all six continuations follow the injected emotion and describe facial cues consistent with that interpretation. These selected cases show coherent changes in generated interpretation.

Consensus vectors mapped into each model

| | Qwen3-VL | InternVL3.5 | Ministral-3 |
|---|---|---|---|
| Anger | ↑ 愤怒, 怒, 强硬, 咆, 毁灭<br>↓ 温馨, 温暖, 愉快, pleasantly, pleasant | ↑ 对抗, 愤怒, 破坏, 强硬, 攻击<br>↓ warmth, 温暖, gentle, 温馨, warm | ↑ aggressive, aggression, -force, force, angry<br>↓ joy, happy, sweet, blush, warm |
| Disgust | ↑ 秽, 恶心, disgusting, 鄙, 不堪<br>↓ 惊喜, joyful, 欣喜, 喜悦, 明亮 | ↑ 毒性, 较差, 这些东西, 毒, 恶心<br>↓ joyful, beautifully, beautiful, 明亮, vibrant | ↑ unpleasant, disgust, dislike, taste, aggressive<br>↓ joy, light, beam, resonant, Frag |
| Fear | ↑ 恐惧, 绝望, Horror, 恐怖, terrified<br>↓ proudly, 愉快, 微笑, 休闲, smiling | ↑ 恐惧, 黑暗, 不安, 害怕, 紧张<br>↓ 微笑, 自豪, charming, sunny, smiling | ↑ frightened, horror, alarm, terrifying, darkness<br>↓ sunshine, smile, smiling, happy, smiles |
| Happiness | ↑ 愉快, 开心, smiles, smiling, happily<br>↓ 绝望, 不堪, 恐惧, 恐怖, 无助 | ↑ smiling, joy, smiles, joyful, smile<br>↓ Worse, 暴力, 苦, 严重, 死亡 | ↑ happy, smile, joy, happiness, smiles<br>↓ unpleasant, alarm, horror, cold, harsh |
| Sadness | ↑ 静静, 黯, sadness, sorrow, 悲伤<br>↓ enthusiastically, enthusiastic, 夸张, hilarious, enthusiasm | ↑ 静静, quiet, 安静, quietly, silence<br>↓ 搞笑, enthusiastically, enthusiastic, 夸张, grinned | ↑ sorrow, sadness, solitude, mourning, ache<br>↓ fun, grin, laughs, entertaining, fren |
| Surprise | ↑ 惊讶, 惊奇, 鸷, 惊, 惊喜<br>↓ 不忍, scars, bitterness, 痛, 苦难 | ↑ 惊讶, surprise, 惊, intriguing, curiosity<br>↓ morale, 不断的, 饱和, 损, scars | ↑ shock, blinking, curiosity, blinked, Shock<br>↓ smokers, strongest, feelings, agner, anonym |

Figure 26: Lexical readouts of the three-model consensus vectors. After ImageNet alignment and averaging over all 21 model–source sets, each vector is mapped back to each model's coordinates. The figure shows the five highest- and lowest-scoring tokens per emotion under that model's lexical readout.

**Consensus vectors retain emotion-related lexical readouts.** We also explore whether the consensus vectors retain emotion-related words in each model's vocabulary. We average the 21 model–source sets in the ImageNet-aligned frame to obtain one consensus vector $\bar{\boldsymbol{v}}_e$ per emotion. We then map each vector back to model $m$ through $\mathbf{R}_m^\top$ and apply that model's lexical readout, as in Appendix C.1.

Figure 26 shows emotion-related promoted tokens across all three models, with some opposite-emotion terms among the suppressed tokens. The pattern also appears in Ministral-3 despite its near-zero native cosine with the Qwen-based models before alignment. The coordinate maps use ImageNet activations without emotion labels, while the consensus averages corresponding emotions after mapping. This lexical evidence complements the projection and steering results.

### F.4 Held-Out Generalization and Model-Size Ablation

The three core models have residual width 4096, and Qwen3-VL and InternVL3.5 use related language backbones. To test generalization beyond the models used to construct the consensus, we evaluate Gemma-3-12B and Gemma-3-27B. Neither variant contributes to consensus construction. Their different residual widths also require cross-width alignment before matched-vector cosine can be computed.

**Held-out correspondence persists across unequal residual widths.** We first explore whether a map fitted on five emotions predicts the sixth across unequal residual widths. Figure 27a,b and Table 27 report both directions between each Gemma-3 variant and the three core models, across seven sources. For 12B, LOEO cosine ranges from 0.16 to 0.52, with median 0.36. All 42 model–source comparisons exceed the permutation threshold ($z = 2.37$–$3.70$), while all 24 replacement controls remain below $z = 1$. For reference, Ministral-3→Qwen3-VL gives 0.13–0.45 with median 0.34 under the same held-out criterion (Table 11).

The 27B model shows a similar pattern. LOEO cosine ranges from 0.13 to 0.54, again with median 0.36, and all 42 comparisons exceed the threshold ($z = 2.40$–$3.78$). Its 24 replacement controls remain below $z = 2$, with cosine between $-0.003$ and 0.013. Held-out correspondence therefore persists at both sizes, without a consistent gain from the larger model.

**Consensus vectors correspond to the held-out Gemma-3 vectors.** We next test the consensus vectors constructed without either Gemma-3 variant. We keep the three-model consensus fixed and learn a cross-width map from paired image activations, without using the six emotion vectors or their

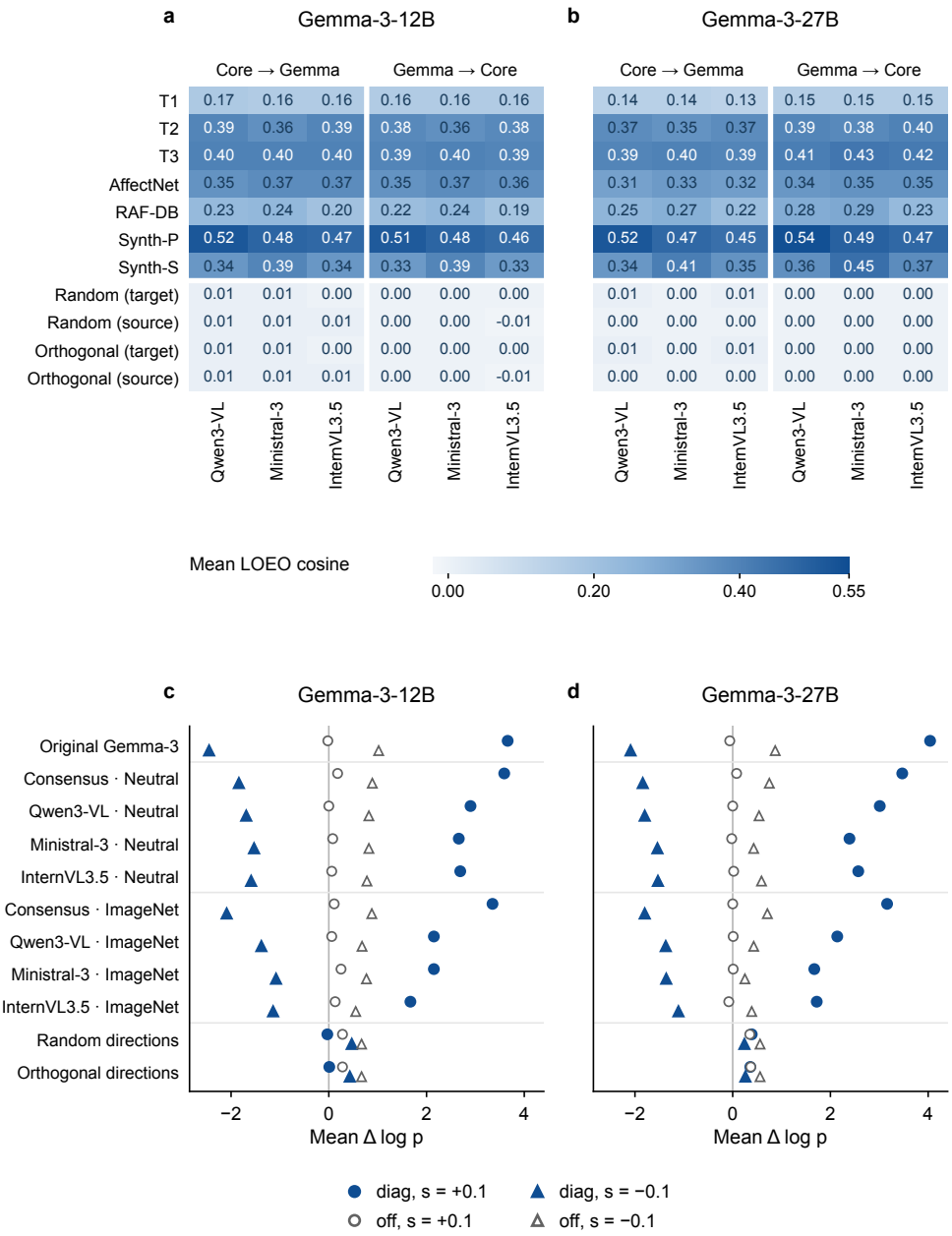



Figure 27: Gemma-3 generalization at 12B and 27B. (a,b) Mean six-emotion LOEO cosine for seven sources and both directions with each core model. The last four rows are RAF-DB replacement controls. (c,d) AffectNet steering at $s = \pm 0.1$. Filled/open markers show diag/off, and circles/triangles show positive/negative injection. Neutral denotes neutral-face mapping. Gemma-3 is excluded from consensus construction. Tables 27 and 28 give exact values.

Table 27: Mean six-emotion LOEO cosine between Gemma-3 and each core model in both directions. Unequal-width fits follow Appendix B.7. RAF-DB controls replace source or target vectors with matched-norm random or orthogonal directions. Source labels follow Table 1.

| | Source model → Gemma-3 | | | Gemma-3 → target model | | |
|---|---|---|---|---|---|---|
| Source / control | Qwen3-VL | Ministral-3 | InternVL3.5 | Qwen3-VL | Ministral-3 | InternVL3.5 |
| *Gemma-3-12B* | | | | | | |
| T1 | 0.17 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 |
| T2 | 0.39 | 0.36 | 0.39 | 0.38 | 0.36 | 0.38 |
| T3 | 0.40 | 0.40 | 0.40 | 0.39 | 0.40 | 0.39 |
| AffectNet | 0.35 | 0.37 | 0.37 | 0.35 | 0.37 | 0.36 |
| RAF-DB | 0.23 | 0.24 | 0.20 | 0.22 | 0.24 | 0.19 |
| Synth-P | 0.52 | 0.48 | 0.47 | 0.51 | 0.48 | 0.46 |
| Synth-S | 0.34 | 0.39 | 0.34 | 0.33 | 0.39 | 0.33 |
| *RAF-DB null controls* | | | | | | |
| Random (target) | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Random (source) | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | −0.01 |
| Orthogonal (target) | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Orthogonal (source) | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | −0.01 |
| *Gemma-3-27B* | | | | | | |
| T1 | 0.14 | 0.14 | 0.13 | 0.15 | 0.15 | 0.15 |
| T2 | 0.37 | 0.35 | 0.37 | 0.39 | 0.38 | 0.40 |
| T3 | 0.39 | 0.40 | 0.39 | 0.41 | 0.43 | 0.42 |
| AffectNet | 0.31 | 0.33 | 0.32 | 0.34 | 0.35 | 0.35 |
| RAF-DB | 0.25 | 0.27 | 0.22 | 0.28 | 0.29 | 0.23 |
| Synth-P | 0.52 | 0.47 | 0.45 | 0.54 | 0.49 | 0.47 |
| Synth-S | 0.34 | 0.41 | 0.35 | 0.36 | 0.45 | 0.37 |
| *RAF-DB null controls* | | | | | | |
| Random (target) | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| Random (source) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Orthogonal (target) | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| Orthogonal (source) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

labels. Let $\mathbf{G}$ be the fixed shared activation template from Equation 17, and let $\widetilde{\mathbf{X}}_{\text{Gemma}}$ contain centered activations from either Gemma-3 variant on the same neutral-face or ImageNet images. We use a thin singular value decomposition (SVD) to compute the map and transfer each consensus vector,

$$\begin{aligned} \mathbf{G}^\top \widetilde{\mathbf{X}}_{\text{Gemma}} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad \mathbf{R}_{\text{Gemma}} = \mathbf{U}\mathbf{V}^\top, \\ \boldsymbol{v}_e^{\text{shared}\to\text{Gemma}} &= \bar{\boldsymbol{v}}_e \mathbf{R}_{\text{Gemma}}, \qquad e \in \mathcal{E}. \end{aligned} \tag{27}$$

Gemma-3 activations are used only for this mapping step, and transferred vectors are normalized before steering.

Table 28 reports mean cosine with Gemma-3's own AffectNet vectors. Neutral-face and ImageNet maps give 0.63 and 0.42 for 12B, and 0.59 and 0.39 for 27B. The consensus therefore retains correspondence with models that did not contribute to its construction.

**Consensus vectors retain causal effects in Gemma-3-12B.** To test whether this correspondence also supports causal transfer, we inject Gemma-3's own vectors, the mapped consensus vectors, and vectors transferred from each core model separately. Figure 27c,d and Table 28 show the comparisons. For 12B at $s = +0.1$, the original vector raises the matched emotion's log-probability by 3.66. Neutral-face and ImageNet consensus vectors give 3.59 and 3.35. Competing-emotion means are much smaller, although the mapped consensus does not suppress them on average. Negative injection lowers the matched emotion and raises competing emotions.

Under each alignment condition, the 12B consensus vectors produce larger positive matched-emotion effects than vectors transferred from any individual core model. For neutral-face alignment, the individual transfers give 2.66–2.90, despite higher cosine with Gemma-3's own vectors (0.70–0.75) than the consensus (0.63). ImageNet alignment shows the same ordering of positive effects.

Table 28: Gemma-3 steering on AffectNet at image-token positions. Cosine compares mapped vectors with Gemma-3's own AffectNet vectors. Consensus construction excludes Gemma-3. Mapping uses paired activations without emotion-vector fitting. diag and off are mean $\Delta \log p$ for the injected emotion and the other five emotions.

| | | $s = -0.1$ | | $s = +0.1$ | |
|---|---|---|---|---|---|
| Injected vector | Cosine | diag | off | diag | off |
| *Gemma-3-12B* | | | | | |
| Original Gemma-3 vector | 1.00 | −2.45 | +1.02 | +3.66 | −0.02 |
| *Neutral-face alignment* | | | | | |
| Consensus vector | 0.63 | −1.84 | +0.89 | +3.59 | +0.18 |
| Qwen3-VL vector | 0.75 | −1.69 | +0.82 | +2.90 | +0.00 |
| Ministral-3 vector | 0.75 | −1.53 | +0.82 | +2.66 | +0.08 |
| InternVL3.5 vector | 0.70 | −1.59 | +0.78 | +2.69 | +0.06 |
| *ImageNet alignment* | | | | | |
| Consensus vector | 0.42 | −2.09 | +0.88 | +3.35 | +0.11 |
| Qwen3-VL vector | 0.46 | −1.38 | +0.68 | +2.15 | +0.06 |
| Ministral-3 vector | 0.50 | −1.08 | +0.77 | +2.15 | +0.25 |
| InternVL3.5 vector | 0.37 | −1.14 | +0.55 | +1.67 | +0.13 |
| *Null controls* | | | | | |
| Random directions | — | +0.47 | +0.67 | −0.03 | +0.28 |
| Orthogonal directions | — | +0.43 | +0.67 | +0.01 | +0.28 |
| *Gemma-3-27B* | | | | | |
| Original Gemma-3 vector | 1.00 | −2.09 | +0.87 | +4.04 | −0.06 |
| *Neutral-face alignment* | | | | | |
| Consensus vector | 0.59 | −1.84 | +0.75 | +3.47 | +0.08 |
| Qwen3-VL vector | 0.73 | −1.80 | +0.54 | +3.01 | +0.00 |
| Ministral-3 vector | 0.72 | −1.54 | +0.43 | +2.39 | −0.02 |
| InternVL3.5 vector | 0.66 | −1.53 | +0.59 | +2.57 | +0.02 |
| *ImageNet alignment* | | | | | |
| Consensus vector | 0.39 | −1.80 | +0.71 | +3.16 | +0.00 |
| Qwen3-VL vector | 0.42 | −1.37 | +0.43 | +2.14 | +0.01 |
| Ministral-3 vector | 0.43 | −1.36 | +0.25 | +1.67 | +0.01 |
| InternVL3.5 vector | 0.34 | −1.11 | +0.39 | +1.72 | −0.08 |
| *Null controls* | | | | | |
| Random directions | — | +0.24 | +0.56 | +0.39 | +0.35 |
| Orthogonal directions | — | +0.26 | +0.56 | +0.36 | +0.37 |

These comparisons again show that similarity to the target model's vector does not fully predict steering strength.

**The model-size ablation preserves the causal-transfer pattern.** We repeat the causal comparison in Gemma-3-27B to test whether the transfer pattern holds at a larger model size. At $s = +0.1$, the original vector gives $(\text{diag}, \text{off}) = (+4.04, -0.06)$, while the consensus gives $(+3.47, +0.08)$ with neutral-face mapping and $(+3.16, +0.00)$ with ImageNet mapping. Both consensus variants again have larger positive matched-emotion effects than the corresponding single-model transfers. Negative injection reverses the matched-emotion effect, while the random and orthogonal controls lack this selective response. Thus, correspondence and causal transfer persist at both sizes. Because width, depth, and the evaluation subsets also differ, this comparison tests robustness across sizes rather than isolating parameter count alone.

# G Limitations, Future Work, and Broader Impact

## G.1 Limitations

Although we find shared structure and transferable causal effects in emotion representations across sources, modalities, and architectures, our study has several limitations. First, the core analyses focus on three approximately 8B-scale VLMs. Although the held-out Gemma-3 experiments extend the analysis to 12B and 27B, generalization to substantially larger models and more diverse model families remains to be tested. Second, we focus on Ekman's six basic emotions because they provide a common emotion set across the image sources used here. The observed organization remains sensitive to stimulus source and may differ for more complex emotions, continuous affective dimensions, or emotion categories defined in different cultural contexts. Finally, this work focuses on understanding how VLMs represent and use emotion concepts. We do not test whether these findings can be used to improve task performance or reduce emotion-related hallucinations.

## G.2 Future Work

In future work, we plan to test whether the relationships among emotion vectors and their causal effects extend to larger VLMs and more diverse model families. We also aim to extend the analysis to video and speech in multimodal large language models (MLLMs). Beyond the six basic emotions, we plan to study more complex emotions, continuous affective labels, and data from different cultural contexts. By comparing a broader range of stimulus sources, we hope to better understand which input properties drive differences in emotion representations. We will also explore whether the emotion vectors and shared subspace can improve emotion understanding or reduce emotion-related hallucinations. To evaluate these applications, we will measure changes in task accuracy and hallucination rates and check for unwanted effects on unrelated tasks.

## G.3 Broader Impact and Ethics

We study internal emotion representations in VLMs, without claiming to infer a person's true emotional state from facial appearance. Our emotion categories define controlled conditions for studying model representations. They do not provide a complete account of human affect. Model behavior can depend on dataset composition, cultural context, and stimulus source. Using these systems in consequential settings could therefore amplify existing biases or encourage overconfident interpretations of human emotion. These steering techniques could also be used to manipulate affect-related outputs. We therefore view them primarily as tools for understanding and auditing multimodal models and encourage careful validation, privacy protection, and human oversight in downstream applications.

# H Ethics Statement

This study explores emotion representations and their causal effects in VLMs using existing datasets and synthetic stimuli. We use existing facial-expression datasets rather than collecting new data from human participants. However, real face images raise privacy concerns even when they come from established research datasets. We therefore do not redistribute AffectNet or RAF-DB images, and access to these data remains subject to the original providers' terms.

Emotion labels provide controlled categories for our experiments, but they do not establish a person's true emotional state or capture the full range of human emotion. Cultural and demographic differences may also affect both dataset labels and model behavior. Our findings therefore should not be taken as evidence that these models can reliably assess individuals in sensitive settings such as mental health, education, or employment. In addition, emotion steering could be used to manipulate model outputs. We discuss these risks and the need for privacy protection, careful validation, and human oversight in Appendix G.3.

## I REPRODUCIBILITY STATEMENT

Appendix B documents the model configurations, stimulus sources and splits, generation prompts, activation extraction, emotion-vector construction, steering protocol, alignment procedures, and similarity measures. Appendices D–F provide detailed results and controls for the main analyses. We will release the code, experiment configurations, stimulus sample lists, synthetic-generation seeds, and extracted emotion vectors described in Appendix B.8. For datasets that cannot be redistributed, we will provide sample identifiers so that researchers with access can reconstruct the experimental subsets.

## J LLM USAGE STATEMENT

We used LLMs to assist with language polishing. In addition, we used generative models to create the synthetic stories, synthetic facial portraits, and synthetic emotion-evoking scenes in CMES. The data sources, generation models, and prompts are documented in Appendices B.2 and B.3.