Geometric decomposition of prompt steering

Fan L. Cheng **Nikolaus Kriegeskorte**

Abstract

Prompting steers large language models (LLMs) and vision–language models (VLMs) without weight updates, but it remains unclear how a change in instruction reshapes internal representations to produce a behavioral effect. We introduce a nested geometric decomposition framework that treats prompting as a transformation of the representational geometry for the content following the prompt. We ask what class of mathematical transformation best explains the effect of prompting by finding the best alignment between representations of the same stimulus set following different prompts. For each prompt pair, we fit a sequence of increasingly expressive stimulus-invariant maps: translation, rigid transformation with uniform-scaling, rigid transformation with axis-wise scaling, affine, and nonlinear transformations. We then test these maps causally by replacing a single layer's prompt-A hidden state for a new set of stimuli with its mapped counterpart and measuring recovery of prompt-B representational geometry and behavior. Across three LLMs, three VLMs, and six text or image datasets varying in style, emotion, scene content, and number, prompts consistently reshape representational geometry toward the instructed task structure. In the cross-validated nested variance decomposition, much of the prompt-induced activation change is explained by shape-preserving maps: translation and rigid transformation with uniform-scaling. The tier profiles reveal model- and task-specific routing strategies, differing in how much transformation classes explain variance and where along the layer hierarchy their contributions emerge. The account generalizes across datasets: a map fitted on one dataset retains a large portion of its in-domain fit on a held-out dataset. Crucially, although translation and rigid transformation tiers already improve behavioral agreement, affine transformation is the first tier to nearly recover target-prompt task geometry and produces corresponding gains in behavioral agreement. This suggests that cross-dimensional linear mixing may be a key contributor to how prompts reorganize representations toward the instructed task structure. Our framework provides a general way to decompose prompt-induced representational change into interpretable geometric components, revealing how a model routes task-relevant structure to produce prompt-driven behavior.

1 Introduction

Instruction prompts can alter model behavior at inference time without updating parameters. For a fixed stimulus, changing the instruction from "describe the object" to "describe the artistic style," for example, can shift which stimulus dimensions become behaviorally relevant and, consequently, which output the model produces. This raises a mechanistic question: how does a change in instruction reorganize internal representations so as to steer downstream behavior?

Two complementary literatures bear on this question. First, representational analyses show that prompts and context restructure internal geometry, producing measurable changes in representational similarity, manifold capacity, trajectory geometry, probing performance, and related diagnostics [Kirsanov et al., 2025, Hosseini et al., 2026, Park et al., 2025, Polo et al., 2026, Gonzalez-Gutierrez and Hovy, 2025, Davidson et al., 2026, Park et al., 2024]. These effects can be depth-dependent,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

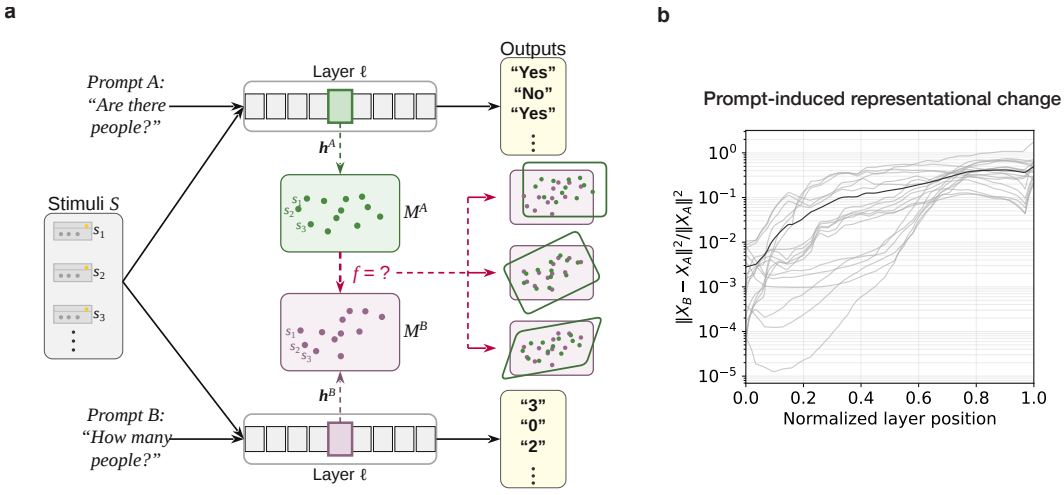

Figure 1: **(a)** Two prompts A ("Are there people?") and B ("How many people?") are presented to the same model together with a stimulus set $\mathcal{S}$. We tap the hidden state at a single transformer layer ℓ (highlighted) to obtain the layer-ℓ manifolds $\mathcal{M}^A = \Phi^A(\mathcal{S})$ (green) and $\mathcal{M}^B = \Phi^B(\mathcal{S})$ (purple). The same forward pass continues past ℓ and yields a prompt-specific output per stimulus ("Yes/No" vs. counts). We ask whether a systematic map $f : \mathcal{M}^A \to \mathcal{M}^B$ exists, how complex it has to be, and whether it causally drives behaviour. **(b)** Prompt-induced representational change as a function of normalised layer depth. Each gray curve is one pair of model and dataset from the three LLMs (OPT-2.7B, Llama3-8B, Qwen3-8B) on three text datasets and three VLMs (BLIP-2, LLaVA-OneVision, Qwen3-VL) on three image datasets; the black curve is their mean. The change is small but non-zero in early layers (10^{-4}–10^{-2}) and grows nearly monotonically with depth, reaching 10^{-1}–10^{0} at the top, with a consistent shape across architectures and modalities.

with some studies suggesting that intermediate layers provide especially informative or compressed representations and later layers more directly support prediction [Jiang et al., 2025, Skean et al., 2024, 2025]. Second, work on activation steering shows that relatively simple activation-space interventions, such as additive vectors, rotations, and affine maps, can modulate prompt-induced behavioral effects [Turner et al., 2023, Zou et al., 2023, Singh et al., 2024, Vu and Nguyen, 2025]. Taken together, these findings suggest that natural prompting may itself be implemented through a low-complexity transformation in activation space.

What remains unclear, however, is the *transformation class* of the prompt-induced change. Existing work on the geometry of prompting has established that prompts alter internal representations, but it has typically characterized those changes through scalar metrics rather than decomposing them into interpretable geometric components. Conversely, activation-steering studies show that engineered low-complexity operators can influence behavior, but they rarely estimate such operators from natural prompting itself or test whether the recovered operators account for target-prompt geometry induced by an instruction change.

We address this gap with a nested geometric decomposition framework. For each model, layer, dataset, and prompt pair, we fit a single stimulus-invariant alignment map from source-prompt activations to target-prompt activations and evaluate it on held-out stimuli. The operator hierarchy separates centroid shifts, global Procrustes alignment, axis-wise scaling in the Procrustes-aligned basis, cross-dimensional feature mixing, and nonlinear residual structure. This hierarchy is motivated by classical Procrustes analysis and by shape-based approaches to comparing representations under explicit transformation classes [Gower, 1975, Williams et al., 2021, Kornblith et al., 2019, Williams, 2024, Harvey et al., 2024, Barbosa et al., 2025]. In parallel, we use RSA and silhouette analyses to test whether prompting reshapes representational geometry toward the instructed task structure; RSA compares representational dissimilarity matrices and is designed to relate representational geometry to candidate task or model structures [Kriegeskorte et al., 2008, Lin and Kriegeskorte, 2024]. Finally, we apply each fitted map as a causal activation intervention, replacing a source-prompt hidden state

with its mapped counterpart and measuring whether the resulting representation and output recover those of the target prompt. The resulting tier profiles reveal the prompt-routing strategy a model uses to follow an instruction: which transformation tiers carry the change, where they emerge across layers, and whether they recover task-aligned geometry and behavior.

Contributions.

1. We introduce a *nested geometric decomposition* of prompt-induced hidden-state change across five transformation classes: translation, rigid transformation with uniform scaling, rigid transformation with axis-wise scaling, affine, and nonlinear, yielding a unique contribution for each tier (§3).
2. We develop a *causal intervention protocol* that applies the fitted map from each tier at a single layer on held-out stimuli and evaluates both representational geometry and behavioural recovery relative to the target prompt (§5).
3. Across LLMs, VLMs, and six datasets, we find that prompting consistently reshapes representational geometry toward the instructed task structure, that much of the prompt effect is captured by low-complexity maps, and that model families, datasets, and prompt-pair groups differ systematically in the relative contribution of transformation tiers and in where these contributions emerge across depth, revealing distinct prompt-routing strategies (§6).

2 Related work

Geometry of prompting and in-context representations. Prior work has characterized the geometric effects of prompting and in-context structure primarily through changes in scalar metrics. A previous study analyzes zero-shot, few-shot, and soft prompting using manifold-capacity tools, showing that different prompting regimes can induce distinct representational mechanisms for task adaptation [Kirsanov et al., 2025]. Other work characterizes how context and in-context examples reorganize internal representations into compact, identifiable substrates that drive in-context generalization, sucha as attention-head circuits, task vectors, and trajectory or phase-transition-like geometry [Hosseini et al., 2026, Park et al., 2025, Yang et al., 2026, Hendel et al., 2023, Todd et al., 2024]. A complementary line documents the dynamic, distributed, and not always task-aligned representational effects of prompting [Gonzalez-Gutierrez and Hovy, 2025, Davidson et al., 2026, Li et al., 2025b, Polo et al., 2026, Li et al., 2025a, Simhi et al., 2026]. Relative to this line of work, our contribution is to estimate explicit transformation classes between prompt-conditioned representation manifolds and to test whether the transformations causally recover both representational geometry and behavior.

Activation steering and geometric interventions. We organize prior steering methods within the transformation hierarchy $\mathcal{F}_{\mathrm{T}} \subset \mathcal{F}_{\mathrm{O_u}} \subset \mathcal{F}_{\mathrm{O_a}} \subset \mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$ (Fig. 3), distinguishing *global* operators, which is stimulus-invariant, from *input-conditioned* interventions, whose effective transformation depends on the current stimuli. Most existing methods are naturally described at the translation tier, where a stimulus-invariant direction is added to the hidden state [Turner et al., 2023, Rimsky et al., 2024, Li et al., 2023b, Zou et al., 2023, Templeton, 2024, Wang et al., 2025, Davidson et al., 2026, Singh et al., 2024]. A smaller group of methods fits a single global linear or affine intervention [Postmus and Abreu, 2024, Sheng et al., 2026, Wu et al., 2024, 2025, Singh et al., 2024]. Rotation-based methods and piecewise-affine optimal-transport edits are also relevant, but they are generally not estimated as Procrustes maps between paired representation clouds, and their target-angle or adaptive variants are often input-conditioned [Vu and Nguyen, 2025, Pham and Nguyen, 2024, Abdullaev et al., 2026, Rodriguez et al., 2025, Scialanga et al., 2025]. The rigid transformation with uniform scaling $\mathcal{F}_{\mathrm{O_u}}$ therefore remains comparatively underexplored as a stimulus-set mapping estimator, while strictly nonlinear interventions appear relatively rare [Zhao et al., 2026, Raval et al., 2026]. Independently of the operator class, steering methods also differ in the stage of the forward pass at which the intervention is applied [Lee et al., 2024, Nguyen et al., 2025, Dang and Ngo, 2026], with a related line operating in weight space rather than activation space [Fierro and Roger, 2025], and recent audits have argued for more systematic geometric and mediator-typology characterizations [Venkatesh and Mahendran Kurapath, 2026, Tan et al., 2024, Im and Li, 2025, Wehner et al., 2025]. In this paper, our goal is not to engineer operators that induce a desired behavior, but to use the same hierarchy as a measurement framework for natural prompting. The finding that low-complexity linear maps recover most of the prompt-induced effect is consistent with the strong empirical performance

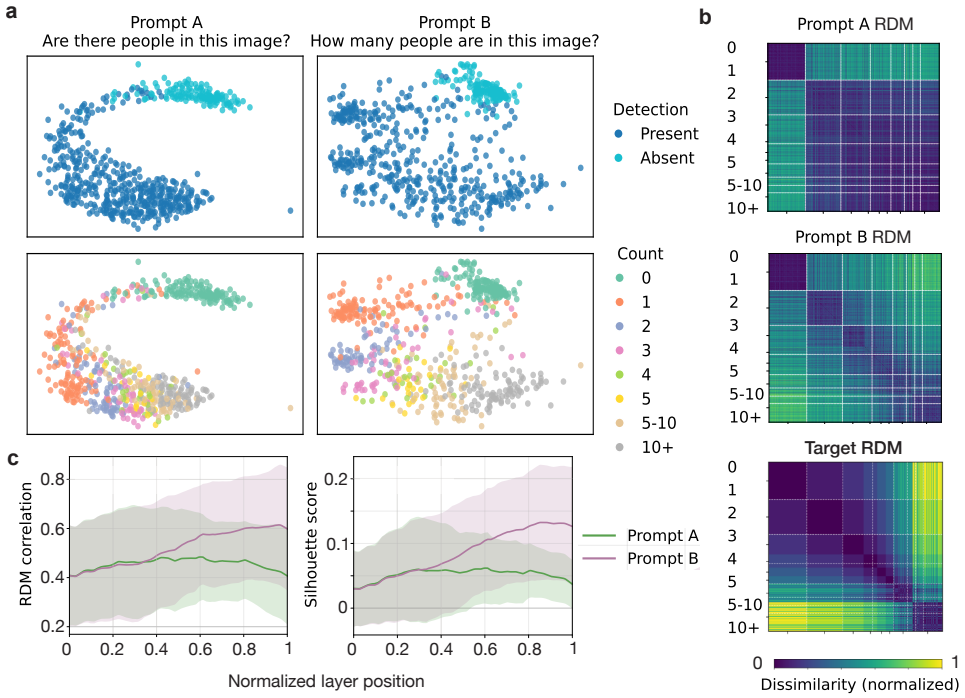

Figure 2: **Prompting reshapes representational geometry toward the instructed task structure.** **(a)** Two-dimensional MDS visualization of the layer-27 representations for LLaVA-OneVision on 1,000 COCO images under two prompts. Each point represents one image; the same MDS coordinates are shown twice and colored by the task-relevant labels for each prompt. **(b)** Representational dissimilarity matrix (RDM) for the same model, layer, and image set as in panel (a). The top two RDMs are computed from the prompt-A and prompt-B hidden states, respectively. The bottom RDM is the target task RDM for prompt B, constructed from numerical distances between count labels. **(c)** Layerwise alignment with the target task structure (pooled across models and datasets). Left: Spearman correlation between each prompt-induced RDM and the corresponding target RDM, computed separately for prompt A and prompt B. Right: silhouette score using the target labels, which measures whether stimuli with the same target label form compact, well-separated clusters in representation space. The shaded regions show ± 1 SD.

of translation- and affine-class steering methods, which may succeed in part because they approximate the activation changes and representational geometry induced by prompting itself.

3 Method: nested geometric decomposition

Preliminaries. Let $\mathcal{S}$ denote the stimulus space (a measurable set of images or text excerpts), and let $\mathcal{P}$ be a finite set of prompts. Let M be a (vision–)language model with L transformer blocks; fix an analysis layer $\ell \in \{1, \dots, L\}$. For each prompt $p \in \mathcal{P}$, the model defines a *feature map*

$$\Phi^p : \mathcal{S} \longrightarrow \mathbb{R}^D, \qquad s \longmapsto \mathbf{h}^p(s), \tag{1}$$

returning the layer-ℓ hidden state at the final input-prompt token. We call its image

$$\mathcal{M}^p := \Phi^p(\mathcal{S}) \subset \mathbb{R}^D$$

the *manifold* of prompt p. The ambient space $\mathbb{R}^D$ carries the canonical Euclidean inner product and the induced Frobenius metric on $\mathbb{R}^{N \times D}$.

Paired observations across prompts. Fix a pair of prompts $A, B \in \mathcal{P}$ and a finite stimulus set $\{s_i\}_{i=1}^N \subset \mathcal{S}$ (indices i, j range over stimuli). For each stimulus i we obtain a paired observation $\big(\Phi^A(s_i),\, \Phi^B(s_i)\big) \in \mathbb{R}^D \times \mathbb{R}^D$, and stack these into representation matrices

$$\mathbf{X}^A,\ \mathbf{X}^B \in \mathbb{R}^{N \times D}, \qquad \big[\mathbf{X}^p\big]_{i,:} = \Phi^p(s_i)^\top. \tag{2}$$

Problem statement. We ask whether there is a map $f : \mathbb{R}^D \to \mathbb{R}^D$ that approximately satisfies $f \circ \Phi^A \approx \Phi^B$ on $\mathcal{S}$ (Fig. 1**a**). We call a fitted map *global* if its parameters are *stimulus-invariant*. We contrast this with *input-conditioned* steering rules, whose effective direction, angle, coefficient, or application mask varies with the current stimuli. We ask how expressive a global map must be before it can explain, and causally reproduce, the representational effect of changing prompts.

Hypothesis classes. We approximate f by elements of an increasing chain of transformation families on $\mathbb{R}^D$, ranging from classical Lie-group actions to more general affine and nonlinear families:

$$\underbrace{\mathbb{R}^D}_{\text{translations } T(D)} \hookrightarrow \underbrace{\mathbb{R}^D \rtimes \big(\mathbb{R}_{>0} \times O(D)\big)}_{\text{similarity group } \mathrm{Sim}(D)} \hookrightarrow \underbrace{\mathbb{R}^D \rtimes \big(\mathrm{Diag}(D) \times O(D)\big)}_{\text{axis-scaled rigid maps}} \hookrightarrow \underbrace{\mathbb{R}^D \rtimes \mathrm{GL}(D)}_{\text{affine group } \mathrm{Aff}(D)} \hookrightarrow \underbrace{\mathcal{F}_{\mathrm{N}}}_{\text{nonlinear maps}} . \tag{3}$$

Each level corresponds to a hypothesis class of mappings on hidden vectors $\mathbf{x} \in \mathbb{R}^{1\times D}$, with free intercept parameters $\mathbf{a}, \mathbf{b} \in \mathbb{R}^D$:

Class $\mathcal{F}_k$	$f_k(\mathbf{x})$	Parameters
$\mathcal{F}_{\mathrm{T}}$: translation	$\mathbf{x} + \mathbf{b}$	$\mathbf{b} \in \mathbb{R}^D$
$\mathcal{F}_{\mathrm{O_u}}$: rigid transformation and uniform scaling	$\mathbf{b} + c\,(\mathbf{x} - \mathbf{a})\,\mathbf{Q}$	$\mathbf{a}, \mathbf{b} \in \mathbb{R}^D,\ \mathbf{Q} \in O(D),\ c \in \mathbb{R}_{>0}$
$\mathcal{F}_{\mathrm{O_a}}$: rigid transformation and axis-wise scaling	$\mathbf{b} + (\mathbf{x} - \mathbf{a})\,\mathbf{Q}\,\mathbf{D}$	$\mathbf{a}, \mathbf{b} \in \mathbb{R}^D,\ \mathbf{Q} \in O(D),\ \mathbf{D} = \mathrm{diag}(d_1, \ldots, d_D)$
$\mathcal{F}_{\mathrm{L}}$: affine transformation	$\mathbf{b} + (\mathbf{x} - \mathbf{a})\,\mathbf{M}$	$\mathbf{a}, \mathbf{b} \in \mathbb{R}^D,\ \mathbf{M} \in \mathbb{R}^{D\times D}$
$\mathcal{F}_{\mathrm{N}}$: nonlinear transformation	$\mathbf{b} + (\mathbf{x} - \mathbf{a})\,\mathbf{M} + h_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{a})$	$\mathbf{a}, \mathbf{b} \in \mathbb{R}^D,\ \boldsymbol{\theta} \in \mathbb{R}^P$

$\mathcal{F}_{\mathrm{N}}$ is instantiated in our experiments as the affine solution plus a single shared multilayer perceptron $h_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^D$ fitted on the affine residual, so that $h_{\boldsymbol{\theta}} = \mathbf{0}$ recovers $\mathcal{F}_{\mathrm{L}}$ and the containment $\mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$ holds at any width (Appendix C); reported ΔR^2_{N} should therefore be read as a lower bound on what an arbitrarily expressive nonlinear map could explain. We use the right-multiplication convention $\mathbf{xM}$ to match the row-major data layout. Appendix B reports a scaling-only map that lies outside this nested chain and isolates the contribution of rotation from that of axis-wise scaling. The five classes form a strictly nested chain $\mathcal{F}_{\mathrm{T}} \subset \mathcal{F}_{\mathrm{O_u}} \subset \mathcal{F}_{\mathrm{O_a}} \subset \mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$; $\mathcal{F}_{\mathrm{O_a}}$ preserves the rigid alignment found by $\mathcal{F}_{\mathrm{O_u}}$ but allows each aligned axis to be independently rescaled, isolating axis-wise gain modulation from arbitrary linear feature mixing. Inclusion identifications are in Appendix A.

Parameter estimation. Each class is fit on the training fold by minimizing the squared Frobenius reconstruction error

$$\widehat{f}_k \;=\; \arg\min_{f \in \mathcal{F}_k} \; \big\| \mathbf{X}^B - f(\mathbf{X}^A) \big\|_F^2, \tag{4}$$

after centring each prompt condition by its training-fold mean. $\mathcal{F}_{\mathrm{T}}$ has a closed-form mean-shift solution; $\mathcal{F}_{\mathrm{O_u}}$ is solved by orthogonal Procrustes, $\mathcal{F}_{\mathrm{O_a}}$ by Procrustes followed by per-axis least squares on the rigidly-aligned features, and $\mathcal{F}_{\mathrm{L}}$ by ridge regression. The nonlinear class $\mathcal{F}_{\mathrm{N}}$ is instantiated as a shared one-hidden-layer MLP fit by stochastic gradient descent on the same Frobenius criterion. All estimators are evaluated under stratified K-fold cross-validation across stimuli. Closed-form derivations, the MLP capacity and optimizer, and the cross-validation procedure are in Appendix A.

Causal intervention. To probe whether the fitted geometric transformation also *causally* reproduces the effect of prompt B, we run model on stimulus s_i under prompt A and replace the layer-ℓ hidden state at the final input token by $\widehat{f}_k\big(\Phi^A(s_i)\big)$ before continuing the autoregressive forward pass. Let $y_k(s_i)$ denote the resulting output token sequence. As a no-fit oracle reference we also include a level that patches the held-out prompt-B representation directly, $\widehat{f}_{\mathrm{prompt_B}}(\Phi^A(s_i)) := \Phi^B(s_i)$, with the surrounding prompt-A context unchanged. This is distinct from running model end-to-end under prompt B: the prompt tokens, the attention pattern up to layer ℓ, and the post-layer-ℓ processing all start from prompt A's context, so $y_{\mathrm{prompt_B}}$ upper-bounds what *any* fitted single-layer replacement can recover from prompt A alone.

4 Evaluation

For each transform $k \in \{\mathrm{T}, \mathrm{O_u}, \mathrm{O_a}, \mathrm{L}, \mathrm{N}\}$ we report three families of measures:

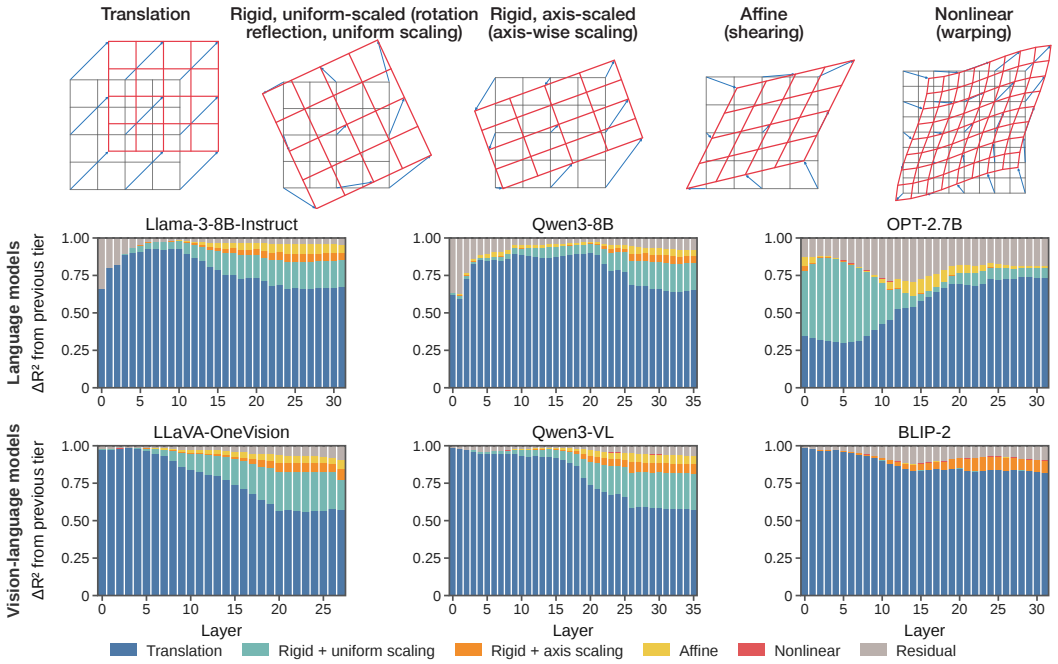

Figure 3: Nested geometric decomposition of prompt-induced representational maps: translation ($\mathcal{F}_{\mathrm{T}}$), rigid transformation with uniform scaling ($\mathcal{F}_{\mathrm{O_u}}$), rigid transformation with axis-wise scaling ($\mathcal{F}_{\mathrm{O_a}}$), affine transformation ($\mathcal{F}_{\mathrm{L}}$), and nonlinear transformation($\mathcal{F}_{\mathrm{N}}$). Top: Schematic of the additional geometric freedom introduced by each tier (listed in parentheses); blue and red grids show the source and transformed representations, respectively. Middle: Incremental explained variance over the preceding tier, ΔR^2, as a function of normalised layer depth for LLMs, pooled across datasets and prompt pairs. Bottom: The same decomposition for VLMs.

1. **Incremental explained variance:** Let $\mathrm{RSS}_k := \left\|\mathbf{X}^B - \widehat{f}_k(\mathbf{X}^A)\right\|_F^2$ on held-out stimuli for $k \in \{\mathrm{T}, \mathrm{O_u}, \mathrm{O_a}, \mathrm{L}, \mathrm{N}\}$, and let $\mathrm{RSS}_0 := \left\|\mathbf{X}^B - \mathbf{X}^A\right\|_F^2$ be the no-transform residual. We define the *cumulative* cross-validated R^2 of tier k as $R^2_k = (\mathrm{RSS}_0 - \mathrm{RSS}_k)/\mathrm{RSS}_0$ and the *incremental* R^2 contribution of tier k over its predecessor $k-1$ in the nested chain as $\Delta R^2_k = R^2_k - R^2_{k-1}$. These increments together with the residual unexplained fraction $R^2_{\mathrm{resid}} = \mathrm{RSS}_{\mathrm{N}}/\mathrm{RSS}_0$ sum to 1 by construction. Individual increments may be negative under cross-validation when a more expressive class generalizes worse on held-out stimuli;
2. **representational geometry metrics:** Spearman correlation between the data RDM of $\widehat{f}_k(\mathbf{X}^A)$ and a category-derived target RDM over $\{s_i\}$ based on the task structure of prompt B, and the silhouette score $s_i = (b_i - a_i)/\max(a_i, b_i)$, where a_i is the mean within-category distance and b_i is the mean distance to the nearest other category;
3. **behavioral recovery:** a per-dataset keyword evaluator scores each intervened output $y_k(s_i)$ on prompt B's ground-truth attribute label (e.g. presence of the target style word, correct numeric count), yielding two metrics: *relevance* (does the text address the target attribute) and *accuracy* (is the answer correct for that stimulus). The oracle-patched output $y_{\mathrm{prompt_B}}(s_i)$ provides the upper-bound reference.

5 Experiments

Models. We evaluate six open-weight transformer models. On the language models: **OPT-2.7B**, **Meta-Llama-3-8B-Instruct**, and **Qwen3-8B**. On the vision language models: **BLIP-2** (OPT-2.7B backbone, Q-Former bridging), **LLaVA-OneVision-7B** (vision encoder/projector with Qwen2 language backbone, multimodal instruction-tuned), and **Qwen3-VL-8B** (native vision-language model with interleaved text-image/video pretraining).

Table 1: Prompt-pair groups. Examples show the prompt pairs for EmotionalStory dataset.

Group	Relation	Pair type	Example (Prompt $A \rightarrow$ Prompt B, EmotionalStory)
G1	Cross-attribute	open $\rightarrow$ open	"What topic is this text about?" $\rightarrow$ "What emotion does this text express?"
G2	Cross-attribute	specific $\rightarrow$ specific	"Is this about a career?" $\rightarrow$ "Does this text express joy?"
G3	Within-attribute	open $\rightarrow$ specific	"What emotion does this text express?" $\rightarrow$ "Does this text express joy?"
G4	Within-attribute	specific $\rightarrow$ specific	"Does this text express joy?" $\rightarrow$ "Does this text express sadness?"
G5	Irrelevant source	irrelevant $\rightarrow$ open	"What is the capital of France?" $\rightarrow$ "What emotion does this text express?"
G6	Irrelevant source	irrelevant $\rightarrow$ specific	"What is the capital of France?" $\rightarrow$ "Does this text express joy?"

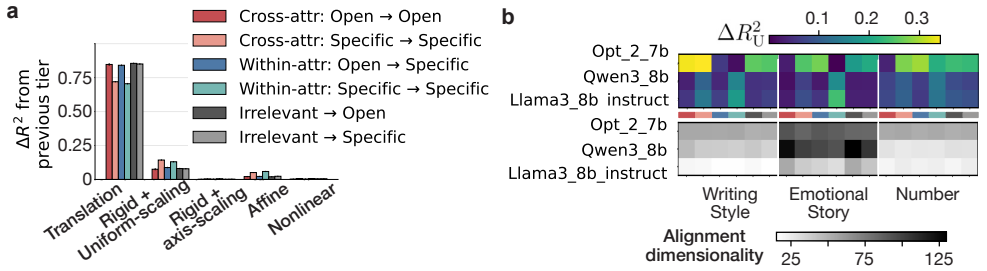

Figure 4: **Comparison of prompt-pair groups and datasets. (a)** Cross-validated ΔR^2, averaged over datasets, models, and layers for each prompt-pair group. Specific prompt-pair groups show distinct decomposition profiles, regardless of whether the paired prompts differ across attributes or within an attribute. **(b) Top:** $\Delta R^2_{\mathrm{O_u}}$, the additional variance explained by rotation/reflection and uniform scaling beyond translation. **Bottom:** alignment dimensionality D. Larger D indicates that more dimensions contribute appreciably to the prompt-pair alignment. OPT-2.7B shows larger and more dataset-dependent $\Delta R^2_{\mathrm{O_u}}$, whereas Llama3-8B shows the lowest alignment dimensionality. EmotionalStory (story topic $\rightarrow$ emotion) shows higher alignment dimensionality across models.

Datasets. The three text datasets (EmotionalStory, WritingStyle, Number) are used by the LLMs; the three image datasets (EmoSet, StyleTransfer, COCO) are used by the VLMs. Each dataset has a two-attribute factorial structure ($\mathrm{attr}_A \times \mathrm{attr}_B$) that the prompt taxonomy exploits for cross- and within-attribute comparisons. Dataset sizes, attribute levels, construction procedures, and prompt templates are in Appendix E. Prompts are written at three levels: *open* (free-form attribute query), *specific* (single-value yes/no query), and *irrelevant* (task-unrelated factual question), and combined into six pair groups (Table 1). G3 and G4 are restricted to the secondary attribute axis; all pairs are evaluated in the forward direction only. See details in Appendix E.

Implementation. We extract hidden states at the last input-prompt token from all transformer layers, and fit and evaluate every transform under five-fold cross-validation. For causal interventions we inject the fitted transform at a single layer and decode greedily up to 50 tokens. Hyperparameters, layer indexing per model family, MLP architecture, and evaluator dictionaries are in Appendix. F.

6 Results

Prompting reshapes representational geometry toward the instructed task structure. We first ask how much change in the hidden activations is induced by prompting. We feed the same stimulus set under two cross-attribute open-ended prompts (G1) and measure the normalised squared Frobenius distance between the prompted hidden states (Fig. 1**b**). The change grows nearly monotonically with depth, which is consistent across other prompt-pair groups (Appendix I).

We then analyzed whether this prompt-induced change carries the geometrical structure required by prompt B (Fig. 2). The example 2D MDS visualization of the prompted hidden states of LLaVA-

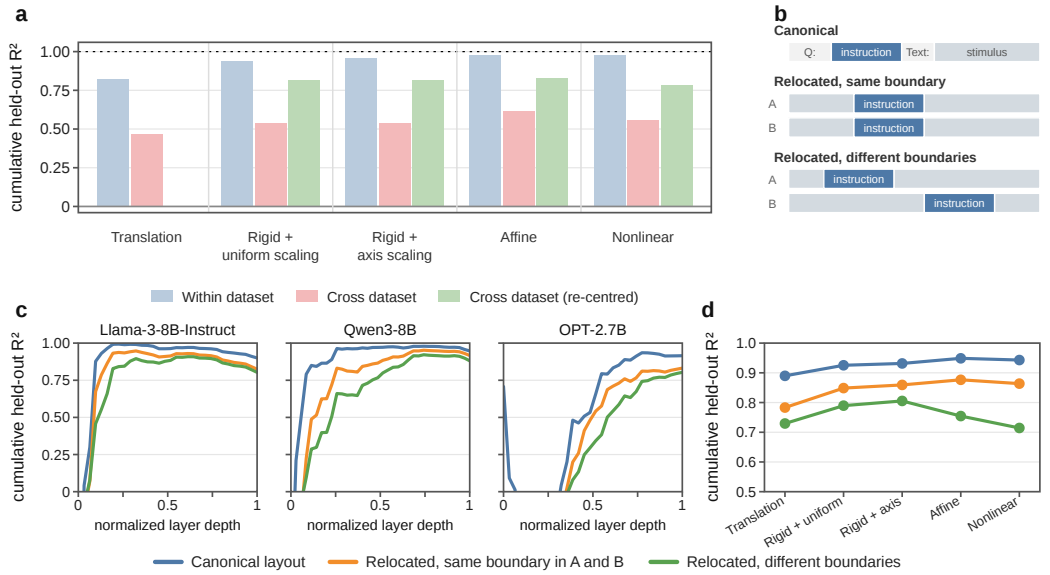

Figure 5: Generalisation across datasets and input layouts. **(a)** Cross-dataset transfer for Llama-3-8B-Instruct and LLaVA-OneVision-7B: per tier, cumulative held-out R^2 within one dataset, on a held-out dataset, and after re-centring the transferred map on the target means. **(b)** The three layouts: the instruction first behind a header, or spliced into an interior sentence boundary, at the same boundary in prompts A and B or at different boundaries. **(c)** Cumulative held-out R^2 at the $\mathcal{F}_{\mathrm{O_a}}$ tier against normalised depth, one panel per model and one curve per layout. **(d)** The same along the tier ladder, averaged over models and over the deepest half of layers; note the restricted vertical range.

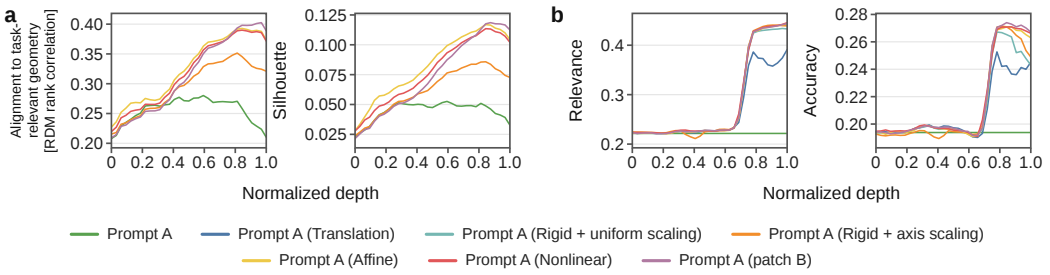

Figure 6: Causal validation of the nested geometric decomposition. At each layer, we replace prompt-A's hidden states with the transformed counterpart for one fitted transformation type and then run the model forward through the remaining layers. **(a)** RDM correlation (left) and silhouette score (right). Axis-scaled rigid transformations improve representational alignment, affine maps produce a larger gain, and nonlinear maps yield the strongest recovery. **(b)** Behavior evaluation on prompt-B's task, measured by Relevance (left) and Accuracy (right). Beyond the translation tier, higher-capacity tiers bring behavioural recovery closer to prompt B.

OneVision-7B for 1000 COCO images demonstrates that the representations of images were separated along the attribute the prompt targets (Fig. 2**a**; see Appendix J for other models and datasets). The representational dissimilariy matrix (RDM)s that computes pair-wise distances between image representations make this explicit (Fig. 2**b**): when stimuli are sorted by count, prompt B's RDM displays the count-graded ordinal block structure of the target count-RDM (bottom panel), while prompt A's does not. The layer-wise alignment with each prompt's target attribute grows monotonically with depth and clearly separates prompt B from prompt A in the top layers (Fig. 2**c**; pooled across models and datasets, see Appendix K for individual results). Higher RDM correlations indicate higher alignment with the task structure of prompt B and higher silhouette scores indicate better separation among categories of the target attribute.

Nested geometric decomposition. To quantify the extent to which prompt-induced representational changes are captured at each level of transformation complexity, we fit the nested family of models $\mathcal{F}_{\mathrm{T}} \subset \mathcal{F}_{\mathrm{O_u}} \subset \mathcal{F}_{\mathrm{O_a}} \subset \mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$ at every layer of each model for prompt pairs across six groups, and compute the *incremental* R^2 contributed by each tier on held-out images (Fig. 3; Fig. 4**a**; see Appendix L for group-specific results). Appendix H shows that the decomposition is unchanged when the stimulus set is enlarged and when the ridge penalty is selected by nested cross-validation, and Appendix D reports the same decomposition fitted in the reverse direction, from prompt B to prompt A, together with the closed-form relations that make the translation tier and part of the rigid tier direction-symmetric. Pure translation accounts for the largest share of explained variance, dominating in early layers and remaining substantial throughout the hierarchy. Higher-complexity tiers contribute additional, smaller increments primarily in mid-to-deep layers, with an overall increasing trend that peaks in the final layers.

Across LLMs, compared with Llama3 and Qwen3, OPT-2.7B relies less on translation in early layers and more on rotation, reflection, scaling, and nonlinear transformations, as well as more on affine transformations in intermediate layers (Fig. 3). This pattern suggests that the models implement distinct early-layer encoding strategies: OPT-2.7B appears to represent prompts as more distributed perturbations that require rotational and higher-dimensional mixing components, whereas Llama3 and Qwen3 may rely on a lower-dimensional "instruction code." Among VLMs, BLIP2 shows a stronger dependence on translation in late layers than LLaVA-OneVision and Qwen3-VL. In BLIP2, multimodal information may be encoded by the Q-Former as a steering vector that persists into deeper layers, whereas LLaVA-OneVision and Qwen3-VL show greater flexibility in information mixing, potentially because of differences in their training protocols. OPT-2.7B is also the only model in this set that received no post-training, which offers one candidate account of its larger and earlier rigid contribution, although the comparison is confounded with architecture, model size, training data and model family (Appendix O).

We additionally compare the explained incremental variance of each tier across different prompt-pair groups (Fig. 4**a**). The five-tier incremental contribution hierarchy ($\mathcal{F}_{\mathrm{T}} > \mathcal{F}_{\mathrm{L}} > \mathcal{F}_{\mathrm{O_u}} > \mathcal{F}_{\mathrm{O_a}} \approx \mathcal{F}_{\mathrm{N}}$) is consistent across prompt-pair groups, indicating that it is not an artifact of any particular pairing regime. Models rely more on rotation, reflection, and affine transformations for paired specific prompts, regardless of whether the pairs are cross-attribute or within-attribute. Appendix M further evaluates generalization across prompt paraphrases and out-of-distribution datasets, showing that fitted transformations are largely stable to semantic rewordings of the target prompt and partially transferable across stimulus distributions, although the relative contributions of transformations remain dataset-dependent.

Generalisation across stimulus distributions and input layouts. Two controls ask whether a fitted map describes the prompt rather than the particular corpus it was fitted on or the particular position the instruction occupies (Fig. 5). Transfer to a held-out dataset is costly at every tier: averaged over Llama-3-8B-Instruct and LLaVA-OneVision-7B and over the six transfer directions, cumulative held-out R^2 falls from 0.98 within a dataset to 0.62 across datasets at the affine tier, and from 0.98 to 0.56 at the nonlinear tier (Fig. 5**a**). Most of that loss is a shift rather than a different map: replacing the source means with the target means lifts the transferred affine fit back to 0.83, and the same correction recovers the nonlinear tier least, leaving it below the affine tier. Appendix N reports the same transfer for all six models and all thirty-six transfer directions. The portable part of a fitted map is therefore its linear part, whereas the extra capacity of the nonlinear tier is the part most tied to the stimulus distribution.

Relocating the instruction preserves the shape of the decomposition and lowers its level (Fig. 5**c**,**d**). Averaged over the three LLMs and over the deepest half of layers, the ladder runs from 0.89 at translation to 0.95 at the affine tier under the canonical layout, from 0.78 to 0.88 when the instruction is relocated to the same sentence boundary in both prompts, and from 0.73 to 0.81 when the two prompts use different boundaries. The tier ordering survives the first relocation. It does not survive the second: with different boundaries the affine and nonlinear tiers fall back below the rigid tiers, which is what fitting a single map across two different instruction positions would predict.

Where rotation matters: explained variance and alignment dimensionality. $\Delta R^2_{\mathrm{O_u}}$ across datasets and prompt-pair groups (Fig. 4**b**, top) shows that OPT-2.7B generally exhibits larger $\Delta R^2_{\mathrm{O_u}}$ than Llama3 and Qwen3. OPT also shows distinct profiles across datasets and groups, with much

higher $\Delta R^2_{\mathrm{O_u}}$ for cross-attribute, open-ended prompt pairs in WritingStyle than in EmotionalStory or Number, which are more translation-dominated (see Appendix L). By contrast, Llama3 and Qwen3 show higher $\Delta R^2_{\mathrm{O_u}}$ for within-attribute, specific prompt pairs than for other groups, whereas OPT shows the opposite trend, relying more on translation for within-attribute, specific prompt pairs.

We further computed the alignment dimensionality, defined as the number of singular values of the centered cross-covariance used by orthogonal Procrustes that exceed $0.01\,\sigma_{\max}$ (Fig. 4**b**, bottom; Appendix S shows the same qualitative pattern under a 0.1 threshold). Higher alignment dimensionality suggests that more dimensions participate in the Procrustes alignment between prompt pairs. EmotionalStory repeatedly shows higher dimensionality than WritingStyle or Number, suggesting that topic-versus-emotion prompt pairs induce a richer, more distributed alignment structure, even when the incremental explained variance is small. Llama3-8B-Instruct tends to have the smallest alignment dimensionalities across models, indicating that its alignment is concentrated in fewer dominant directions than that of OPT-2.7B or Qwen3-8B. The dimensionality of rotation and reflection therefore varies with both task structure and model family.

Causal validation: representational geometry and behavior. We performed an interventional analysis (§3) and computed alignment metrics on held-out stimuli to test whether the fitted transformations can causally recover the representational geometry and behaviour induced by prompt B (Fig. 6; pooled across models, datasets, and prompt-pair group G1). For the interventional analysis, we focus on the open-ended prompt B (G1 and G5) because they preserve non-degenerate multi-label structure for RSA, avoid yes-bias artefacts in behavioural evaluation, and provide a matched comparison between task-relevant and task-irrelevant source prompts.

Translation and uniform-scaled rigid transformations are omitted from Fig. 6**a** because they leave distance-based RDM rankings and silhouette scores unchanged (App. K for additional results). Axis-scaled rigid transformations improve geometric alignment but remain clearly below affine and nonlinear transformations, especially in later layers. This indicates that allowing each Procrustes-aligned axis to be rescaled independently captures part of the prompt-pair transformation, but does not fully account for the representational reorganisation. The additional improvement from affine maps suggests that feature mixing is important for matching prompt-B geometry. The nonlinear map gives the strongest recovery, but its advantage over the affine map is smaller than the gain from the axis-scaled tier to the affine tier. Downstream behavior shows a sharper divide between translation and the higher-capacity transformations (Fig. 6**b**; see Appendix T for text outputs, and Appendix G for an audit of the scorer against a stronger judge). While the translation tier can already substantially steer behaviour, the higher-capacity tiers bring behavioural performance closer to prompt B.

7 Conclusion

Our paper recast prompting as a low-complexity geometric transformation: a translation that does most of the work, rotation and shearing fixes the residual structured component. The effect of the prompts studied here, thus, can be understood as effecting an affine transform of the representation of the content following the prompt. Allowing a nonlinear transform did not explain either the resulting geometry or the resulting behavior substantially better. We see this as evidence that the hidden-state effects of instruction prompting are well approximated by an affine map at the population level, and that the residual fine structure — the rotation, and the small nonlinear part — is where the operational consequences of prompting are disproportionately concentrated. The decomposition is fitted at a single readout position. Appendix P shows that the same low-complexity maps describe every token of the stimulus span, whereas the readout position is the hardest to fit, and Appendices Q and R locate where the instruction is written and where it is read.

Limitations and broader impacts. Our analysis is limited to zero-shot instruction prompts and does not evaluate few-shot demonstrations, soft prompts, or longer-context conditioning. Our intervention replaces the residual stream at a single layer and final input-prompt token; multi-layer or multi-token interventions and more systematic tests under input-distribution shift remain future work. The proposed framework provides an interpretability benefit by decomposing prompt-induced representational change into explicit transformation classes and quantifying their contributions to variance, geometry, and behavior. Such measurements could also inform more efficient activation-level steering. We therefore view the method both as a diagnostic tool for transparency and controlled evaluation, and as a potential basis for deployment-oriented steering techniques.

Acknowledgments and Disclosure of Funding

Funding. Research reported in this publication was supported in part by the National Institute on Drug Abuse of the National Institutes of Health under award number 1R01DA060742. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests. The authors declare no competing interests.

References

Laziz U Abdullaev, Noelle YL Wong, Ryan TZ Lee, Shiqi Jiang, Khoi NM Nguyen, and Tan M Nguyen. Concept heterogeneity-aware representation steering. *arXiv preprint arXiv:2603.02237*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Joao Barbosa, Amin Nejatbakhsh, Lyndon Duong, Sarah E Harvey, Scott L Brincat, Markus Siegel, Earl K Miller, and Alex H Williams. Quantifying differences in neural population activity with shape metrics. *bioRxiv*, pages 2025–01, 2025.

Tal Boger and Chaz Firestone. The psychophysics of style. *Nature Human Behaviour*, pages 1–13, 2025.

Quy-Anh Dang and Chris Ngo. Selective steering: Norm-preserving control through discriminative layer selection. *arXiv preprint arXiv:2601.19375*, 2026.

Guy Davidson, Todd M. Gureckis, Brenden Lake, and Adina Williams. Do different prompting methods yield a common task representation in language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=fy5InEg0OL`.

Constanza Fierro and Fabien Roger. Steering language models with weight arithmetic. *arXiv preprint arXiv:2511.05408*, 2025.

Cesar Gonzalez-Gutierrez and Dirk Hovy. Do prompts reshape representations? an empirical study of prompting effects on embeddings. *arXiv preprint arXiv:2510.19694*, 2025.

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sarah E Harvey, David Lipshutz, and Alex H Williams. What representational similarity measures imply about decodable information. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL `https://openreview.net/forum?id=hqfzH6GCYj`.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, 2023.

Eghbal A Hosseini, Yuxuan Li, Yasaman Bahri, Declan Campbell, and Andrew Kyle Lampinen. Context structure reshapes the representational geometry of language models. *arXiv preprint arXiv:2601.22364*, 2026.

Zhimin Hu, Lanhao Niu, and Sashank Varma. The representational geometry of number. *arXiv preprint arXiv:2602.06843*, 2026.

Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*, 2025.

Jiachen Jiang, Yuxin Dong, Jinxin Zhou, and Zhihui Zhu. From compression to expression: A layerwise analysis of in-context learning. *arXiv preprint arXiv:2505.17322*, 2025.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. *arXiv preprint arXiv:2502.08009*, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Bo Li, Guanzhi Deng, Ronghao Chen, Junrong Yue, Shuo Zhang, Qinghua Zhao, Linqi Song, and Lijie Wen. Rema: A unified reasoning manifold framework for interpreting large language model. *arXiv preprint arXiv:2509.22518*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530, 2023b.

Yuxuan Li, Declan Campbell, Stephanie CY Chan, and Andrew Kyle Lampinen. Just-in-time and distributed task representations in language models. *arXiv preprint arXiv:2509.04466*, 2025b.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Dung V Nguyen, Hieu M Vu, Nhi Y Pham, Lei Zhang, and Tan M Nguyen. Activation steering with a feedback controller. *arXiv preprint arXiv:2510.04309*, 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=pXlmOmlHJZ`.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Van-Cuong Pham and Thien Huu Nguyen. Householder pseudo-rotation: A novel approach to activation editing in llms with direction-magnitude perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13737–13751, 2024.

Alexandre Polo, Chanwoo Chun, and SueYeon Chung. Emergent manifold separability during reasoning in large language models. *arXiv preprint arXiv:2602.20338*, 2026.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*, 2024.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. *arXiv preprint arXiv:2603.09313*, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, marco cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=l2zFn6TIQi`.

Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. Sake: Steering activations for knowledge editing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15966–15978, 2025.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=1vvbzAqdTe`.

Adi Simhi, Fazl Barez, Martin Tutek, Yonatan Belinkov, and Shay B Cohen. Old habits die hard: How conversational history geometrically traps llms. *arXiv preprint arXiv:2603.03308*, 2026.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=WGXb7UdvTX`.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212, 2024.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Eric Todd, Millicent Li, Arnab Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*. ICLR, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large lan-guage models. *arXiv e-prints*, pages arXiv–2602, 2026.

Hieu M. Vu and Tan Minh Nguyen. Angular steering: Behavior control via rotation in activation space. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025. URL `https://openreview.net/forum?id=GU2UeVZrSw`.

Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. Improving llm reasoning through interpretable role-playing steering. *arXiv preprint arXiv:2506.07335*, 2025.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*, 2025.

Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL https://openreview.net/forum?id=zMdnnFasgC.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=L9JM-pxQOl.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fykjplMc0V.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D Manning, and Christopher Potts. Improved representation steering for language models. *arXiv preprint arXiv:2505.20809*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Haolin Yang, Hakaze Cho, Yiqiao Zhong, and Naoya Inoue. Unifying attention heads and task vectors via hidden state geometry in in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FIfjDqjV0B.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek Abdelzaher, Yejin Choi, Manling Li, et al. Odesteer: A unified ode-based steering framework for llm alignment. *arXiv preprint arXiv:2602.17560*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Methods: Nested Geometric Decomposition

Nested-chain inclusions. $\mathcal{F}_{\mathrm{T}}$ is recovered from $\mathcal{F}_{\mathrm{O_u}}$ at $(\mathbf{a}, c, \mathbf{Q}) = (\mathbf{0}, 1, \mathbf{I})$; $\mathcal{F}_{\mathrm{O_u}}$ from $\mathcal{F}_{\mathrm{O_a}}$ at $\mathbf{D} = c\,\mathbf{I}$; $\mathcal{F}_{\mathrm{O_a}}$ from $\mathcal{F}_{\mathrm{L}}$ at $\mathbf{M} = \mathbf{Q}\,\mathbf{D}$ with $\mathbf{Q} \in O(D)$ and $\mathbf{D}$ diagonal; and $\mathcal{F}_{\mathrm{L}}$ from $\mathcal{F}_{\mathrm{N}}$ when $g_{\boldsymbol{\theta}}$ implements a linear map.

Parameter estimation. The four parametric tiers all operate on the centred matrices $\widetilde{\mathbf{X}}^A = \mathbf{X}^A - \boldsymbol{\mu}^A$ and $\widetilde{\mathbf{X}}^B = \mathbf{X}^B - \boldsymbol{\mu}^B$. smallskip *Translation ($\mathcal{F}_{\mathrm{T}}$).* The mean-shift solution is

$$\widehat{\mathbf{b}} \;=\; \boldsymbol{\mu}^B - \boldsymbol{\mu}^A. \tag{5}$$

Uniform-scaled rigid ($\mathcal{F}_{\mathrm{O_u}}$). Compute the cross-covariance $\mathbf{C} = \widetilde{\mathbf{X}}^{A\top}\widetilde{\mathbf{X}}^B$ and its SVD $\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. The orthogonal-Procrustes solution is

$$\widehat{\mathbf{Q}} \;=\; \mathbf{U}\mathbf{V}^\top, \qquad \widehat{c} \;=\; \frac{\langle \widetilde{\mathbf{X}}^A\widehat{\mathbf{Q}},\, \widetilde{\mathbf{X}}^B\rangle_F}{\|\widetilde{\mathbf{X}}^A\|_F^2}. \tag{6}$$

Axis-scaled rigid ($\mathcal{F}_{\mathrm{O_a}}$). Re-use $\widehat{\mathbf{Q}}$ from the $\mathcal{F}_{\mathrm{O_u}}$ fit, form the rigidly-aligned features $\mathbf{Z} = \widetilde{\mathbf{X}}^A\widehat{\mathbf{Q}}$, and fit each diagonal entry by per-axis least squares,

$$\widehat{d}_j \;=\; \frac{\langle \mathbf{Z}_{:,j},\, \widetilde{\mathbf{X}}^B_{:,j}\rangle}{\|\mathbf{Z}_{:,j}\|_2^2}, \qquad j = 1, \dots, D. \tag{7}$$

Affine ($\mathcal{F}_{\mathrm{L}}$). Ridge regression gives the closed form

$$\widehat{\mathbf{M}} \;=\; \big(\widetilde{\mathbf{X}}^{A\top}\widetilde{\mathbf{X}}^A + \lambda\mathbf{I}\big)^{-1}\,\widetilde{\mathbf{X}}^{A\top}\widetilde{\mathbf{X}}^B, \qquad \lambda = 1. \tag{8}$$

$g_{\boldsymbol{\theta}}$ is a two-layer MLP with one hidden layer of size $H{=}512$ and GELU activation. It is trained for 200 epochs over the training-fold rows with mini-batch size $\min(256, N_{\text{train}})$, using AdamW (learning rate 10^{-3}, weight decay 10^{-3}) on the mean-squared reconstruction error (equivalent to the Frobenius criterion up to a $1/(N_{\text{train}}\,D)$ normalization). The seed is set per fold for reproducibility.

For $k \in \{\mathrm{O_u}, \mathrm{O_a}, \mathrm{L}\}$ the empirical risk is invariant under $(\mathbf{a}, \mathbf{b}) \mapsto (\mathbf{a} + \mathbf{v}, \mathbf{b} + f_k(\mathbf{v}) - f_k(\mathbf{0}))$; we adopt the canonical choice $\widehat{\mathbf{a}} = \boldsymbol{\mu}^A$ and $\widehat{\mathbf{b}} = \boldsymbol{\mu}^B$, reducing the remaining estimation to a problem on the centred matrices $\widetilde{\mathbf{X}}^A, \widetilde{\mathbf{X}}^B$. For $\mathcal{F}_{\mathrm{N}}$ we use the same centring convention, although the risk is not invariant under arbitrary translations due to the nonlinearity of $g_{\boldsymbol{\theta}}$.

Cross-validation. We use stratified five-fold cross-validation across each stimuli set; per-fold estimators are fit on the training partition and evaluated on the held-out partition. Cumulative R_k^2 and incremental ΔR_k^2 are computed from held-out residual sums of squares.

Hypothesis interpretation. Each pure transformation class produces a characteristic decomposition profile, summarized in Table 2.

Table 2: Decomposition signatures of pure transformation classes.

Pure transformation	Dominant component	Interpretation
Mean shift $\mathbf{b}$	$\Delta R^2_{\mathrm{T}} \to 1$	concept-vector / steering hypothesis
Rigid + uniform scaling	$\Delta R^2_{\mathrm{O_u}} \to 1$	readout realignment / shape-preserving
Rigid + axis-wise scaling	$\Delta R^2_{\mathrm{O_a}} \to 1$	per-axis gain modulation
General linear map	$\Delta R^2_{\mathrm{L}} \to 1$	arbitrary linear feature mixing
Nonlinear map	$\Delta R^2_{\mathrm{N}} > 0$	genuine nonlinear restructuring
Residual / noise	$R^2_{\mathrm{resid}} > 0$	stimuli-dependent / beyond-MLP capacity

B Axis-wise scaling without rotation

The tier $\mathcal{F}_{\mathrm{O_a}}$ of Section 3 introduces axis-wise scaling and a rigid transformation at the same time, so its increment over $\mathcal{F}_{\mathrm{O_u}}$ does not by itself say which of the two ingredients carries the prompt-induced change. To separate them we additionally fit a scaling-only family that omits the rotation entirely,

$$f_{\mathrm{S_a}}(\mathbf{x}) \;=\; \mathbf{b} + (\mathbf{x} - \mathbf{a})\,\mathbf{D}, \qquad \mathbf{a}, \mathbf{b} \in \mathbb{R}^D, \quad \mathbf{D} = \mathrm{diag}(d_1, \ldots, d_D), \tag{9}$$

estimated by per-axis least squares on the centred matrices under the same centring convention, the same fold construction, the same direction-symmetric RSS_0, and the same cumulative R^2 definition as every other tier.

This family sits deliberately outside the nested decomposition. It contains $\mathcal{F}_{\mathrm{T}}$ at $\mathbf{D} = \mathbf{I}$ and is contained in $\mathcal{F}_{\mathrm{O_a}}$ at $\mathbf{Q} = \mathbf{I}$, so it lies on a second branch between the translation tier and the axis-scaled rigid tier rather than on the main chain. It is also incomparable with $\mathcal{F}_{\mathrm{O_u}}$ in both directions: a rotation with off-diagonal structure is not diagonal, whereas a diagonal matrix with unequal entries is not a positive multiple of an orthogonal matrix. Because neither family nests inside the other, the gap between their cumulative R^2 values is not an incremental variance component and must not be read as one of the ΔR^2 terms that partition the total. We therefore report the scaling-only family only as a cumulative value, placed beside $\mathcal{F}_{\mathrm{O_u}}$, which supplies the complementary ingredient of one rotation together with a single global scale.

Read this way, the comparison against uniform scaling with a rigid transformation suggests that rotation likely contributes more to the prompt-induced representational change than axis-wise scaling does. We state this as a suggestion rather than a claim. The two families differ in parameter count as well as in geometry, and the scaling-only family is not part of the variance partition reported in the main text, so it supports a qualitative ordering of the two ingredients and nothing stronger.

C Residual parameterization of the nonlinear tier

In the submitted implementation the nonlinear tier was a one-hidden-layer MLP of width 512 applied directly to the centred input, whereas the activation dimension D of the studied models ranged from 2560 to 4096. Such a bottleneck cannot represent an arbitrary affine map, so the implemented class did not satisfy the containment $\mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$ that the nested decomposition requires. We therefore parameterize the nonlinear tier as a residual extension of the fitted affine map,

$$f_{\mathrm{N}}(\mathbf{x}) \;=\; \mathbf{b} + (\mathbf{x} - \mathbf{a})\,\mathbf{M} + h_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{a}), \tag{10}$$

where $\mathbf{M}$ is the affine-tier ridge solution of Appendix A and $h_{\boldsymbol{\theta}}$ is the same one-hidden-layer MLP, now trained on the affine residuals. Setting $h_{\boldsymbol{\theta}} = 0$ recovers the affine map exactly, so $\mathcal{F}_{\mathrm{L}} \subset \mathcal{F}_{\mathrm{N}}$ holds by construction at any width of $h_{\boldsymbol{\theta}}$. The nonlinear increment then measures residual nonlinear structure rather than a capacity limitation.

Table 3: Mean cumulative held-out R^2 of the nonlinear tier across all pairs of model and dataset, under three implementations.

Implementation of $\mathcal{F}_{\mathrm{N}}$	Mean cumulative R^2
Submitted: direct MLP of width 512	0.922
Corrected: residual MLP of width 512	0.940
Control: direct MLP widened to the full activation dimension	0.943

We refit all three implementations for every pair of model and dataset (Table 3) and regenerated the causal interventions in both directions with the corrected map. The correction changes the fit only slightly because the affine tier already reaches cumulative held-out $R^2 \geq 0.96$ in 14 of the 18 model-by-dataset pairs, leaving little headroom for any nonlinear refinement. The transformation hierarchy and the conclusions of the main text are unchanged.

D The reverse direction: fitting prompt B to prompt A

All fits reported in the main text estimate a map from prompt A to prompt B. Here we fit the reverse map, from prompt B to prompt A, re-estimating every tier independently in the reverse direction

under the same five-fold cross-validation across stimuli and the same direction-symmetric baseline $\mathrm{RSS}_0 = \|\mathbf{X}^B - \mathbf{X}^A\|_F^2$. Two of the five tiers are constrained analytically by the forward fit, whereas the remaining three are not. We first make that statement precise and then report the measurements.

D.1 Translation is direction-symmetric by construction

For $\mathcal{F}_{\mathrm{T}}$ the mean-shift solution of Appendix A gives, on any training fold,

$$\widehat{\mathbf{b}}_{A\to B} \;=\; \boldsymbol{\mu}^B - \boldsymbol{\mu}^A, \qquad \widehat{\mathbf{b}}_{B\to A} \;=\; \boldsymbol{\mu}^A - \boldsymbol{\mu}^B \;=\; -\,\widehat{\mathbf{b}}_{A\to B}, \tag{11}$$

so the fitted offset for the reverse direction is exactly the negative of the fitted offset for the forward direction. The held-out residuals therefore satisfy

$$\mathbf{X}^A - \big(\mathbf{X}^B + \widehat{\mathbf{b}}_{B\to A}\big) \;=\; -\Big(\mathbf{X}^B - \big(\mathbf{X}^A + \widehat{\mathbf{b}}_{A\to B}\big)\Big), \tag{12}$$

that is, the two residual matrices are negatives of each other. Their squared Frobenius norms coincide, $\mathrm{RSS}_{\mathrm{T},B\to A} = \mathrm{RSS}_{\mathrm{T},A\to B}$, and since RSS_0 is itself direction-symmetric, the held-out cumulative R^2_{T} matches fold by fold, not merely on average. The translation column of Table 4 is identical in the two directions for this reason and carries no information about directional asymmetry.

D.2 Closed-form relations for the uniform-scaled rigid tier

For $\mathcal{F}_{\mathrm{O_u}}$, let $\mathbf{C} = \widetilde{\mathbf{X}}^{A\,\top}\widetilde{\mathbf{X}}^B = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the cross-covariance and its SVD on a training fold. Reversing the direction transposes the cross-covariance, which exchanges $\mathbf{U}$ and $\mathbf{V}$ and leaves $\boldsymbol{\Sigma}$ unchanged, so the two orthogonal-Procrustes solutions are transposes of one another,

$$\widehat{\mathbf{Q}}_{B\to A} \;=\; \widehat{\mathbf{Q}}_{A\to B}^{\top}. \tag{13}$$

The optimal rotation is thus shared. The scale is not: because $\big\langle \widetilde{\mathbf{X}}^A\widehat{\mathbf{Q}},\, \widetilde{\mathbf{X}}^B\big\rangle_F = \mathrm{tr}(\boldsymbol{\Sigma})$, the two scale estimates differ only in their denominator,

$$\widehat{c}_{A\to B} \;=\; \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\big\|\widetilde{\mathbf{X}}^A\big\|_F^2}, \qquad \widehat{c}_{B\to A} \;=\; \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\big\|\widetilde{\mathbf{X}}^B\big\|_F^2}, \tag{14}$$

and both are governed by the single shared alignment coefficient

$$\rho \;\coloneqq\; \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\big\|\widetilde{\mathbf{X}}^A\big\|_F\,\big\|\widetilde{\mathbf{X}}^B\big\|_F} \;\in\; [0,1], \tag{15}$$

which is invariant under exchanging the two prompts and satisfies $\widehat{c}_{A\to B}\,\widehat{c}_{B\to A} = \rho^2$. Substituting the two scale estimates into the held-out residual sums of squares yields the difference identity

$$R^2_{\mathrm{O_u},B\to A} - R^2_{\mathrm{O_u},A\to B} \;=\; \frac{\Big(\big\|\widetilde{\mathbf{X}}^B\big\|_F^2 - \big\|\widetilde{\mathbf{X}}^A\big\|_F^2\Big)\big(1-\rho^2\big)}{\mathrm{RSS}_0}. \tag{16}$$

Equation (16) factorizes the directional gap into a term that depends only on the difference in the dispersion of the two prompt conditions and a term $1-\rho^2$ that measures how far the two centred configurations are from being exactly related by a scaled rotation. Two consequences follow. The gap vanishes when the two conditions have equal dispersion, and, more importantly, it is suppressed by the factor $1-\rho^2$ whenever the rigid tier fits well: directional differences shrink as the fit improves and concentrate in the weakest-fitting pairs.

D.3 The higher tiers are not constrained by the forward fit

No analogous identity exists above $\mathcal{F}_{\mathrm{O_u}}$. The axis-scaled rigid family is not closed under inversion. With the right-multiplication convention of Section 3 the forward map is $f_{\mathrm{O_a}}(\mathbf{x}) = \mathbf{b} + (\mathbf{x}-\mathbf{a})\,\mathbf{Q}\,\mathbf{D}$, whose exact inverse is

$$f_{\mathrm{O_a}}^{-1}(\mathbf{y}) \;=\; \mathbf{a} + (\mathbf{y}-\mathbf{b})\,\mathbf{D}^{-1}\mathbf{Q}^\top, \tag{17}$$

which applies the same two operations in the reverse order: it rescales the axes before rotating, whereas every member of $\mathcal{F}_{\mathrm{O_a}}$ rotates before rescaling. Since $\mathbf{D}^{-1}\mathbf{Q}^\top$ is in general not of the form

$\mathbf{Q}'\mathbf{D}'$ with $\mathbf{Q}' \in O(D)$ and $\mathbf{D}'$ diagonal, the inverse of the forward fit is not itself an admissible reverse-direction hypothesis. A reverse fit optimized independently within $\mathcal{F}_{\mathrm{O_a}}$ is therefore not determined by the forward fit and need not attain the same held-out quality, in either direction of the inequality. The same holds for $\mathcal{F}_{\mathrm{L}}$ and $\mathcal{F}_{\mathrm{N}}$, where the estimator is a regression of one condition on the other and is genuinely direction-asymmetric: the ridge solution $\widehat{\mathbf{M}}$ depends on the second moment of the predictor condition alone, so exchanging predictor and target changes the regularized problem rather than inverting it.

D.4 Measured reverse-direction performance

Table 4: Reverse-direction fits. Mean held-out cumulative R^2 over the deepest 50% of layers, averaged across the three datasets per model. The value shown is $B \to A$, with $A \to B$ in parentheses.

Model	Translation	Uniform scaling + rigid	Axis scaling + rigid	Affine	Nonlinear
BLIP-2	0.852 (0.852)	0.926 (0.951)	0.934 (0.955)	0.947 (0.969)	0.945 (0.968)
LLaVA-OneVision	0.772 (0.772)	0.948 (0.929)	0.964 (0.949)	0.977 (0.967)	0.976 (0.967)
Qwen3-VL	0.820 (0.820)	0.954 (0.942)	0.978 (0.962)	0.990 (0.980)	0.989 (0.978)
Llama-3-8B	0.874 (0.874)	0.955 (0.945)	0.963 (0.965)	0.979 (0.986)	0.976 (0.984)
Qwen3-8B	0.922 (0.922)	0.965 (0.969)	0.970 (0.977)	0.976 (0.983)	0.974 (0.982)
OPT-2.7B	0.749 (0.749)	0.805 (0.827)	0.807 (0.832)	0.847 (0.874)	0.839 (0.866)

The translation column is identical in the two directions for all six models, exactly as the mean-shift argument above requires. Above that tier the qualitative picture is unchanged by reversing the direction: the tiers remain ordered, the affine tier is the best-performing tier in both directions for every model, and the nonlinear tier sits marginally below it. Where the two directions do differ, the discrepancy is largest for OPT-2.7B and BLIP-2, the two models with the lowest cumulative R^2 overall, and smallest for Qwen3-8B and Llama-3-8B, which fit best; this is the pattern that Equation (16) anticipates, and the sign of the difference is not consistent across models.

E Datasets and prompt design

Dataset design. **EmotionalStory** (N=1,920) is a corpus of 8 emotions×8 topics×30 short stories, with name-pool and scenario-pool diversity injections. **WritingStyle** (N=1,440) is a synthetically generated corpus of 6 styles × 4 topics × 60 short passages; style and content vary factorially by rendering the same neutral content seed in each of the six styles. **Number** (N=1,247) uses a mix of pseudo-stimuli, real Wikipedia sentences, and task-template stimuli spanning five cognitive-task framings, with set-property prompts (e.g. "Are there any prime numbers?") that require scanning all numerical tokens rather than matching a single value. Tables 5–7 show representative stimuli from the three text datasets. **EmoSet** (N=1,600) is a balanced 8 emotions × 4 content categories (person, animal, nature, object) subset of the EmoSet benchmark [Yang et al., 2023]. **StyleTransfer** (N=1,920; Boger and Firestone, 2025) consists of photographs rendered in seven styles (six artistic styles plus the original photographs) across four scenes (beach, bedroom, library, mountain), with multiple images per style×scene cell. **COCO** (N=1,000) is a subset of COCO val2017 [Lin et al., 2014] selected for high multi-supercategory coverage ($\geq$ 3 of 12 supercategories per image), probing people-detection and people-count prompts.

Diversity controls in synthetic LLM stimuli. Without explicit controls, the LLM seed generator collapses on a few canonical patterns (e.g. defaulting to "Maya" for joy stories and "Marcus" for anger stories, or repeating one factual seed across all *nature* passages). To prevent this we (i) curate a balanced 100-name pool for EmotionalStory and assign names by a coprime stride so each emotion uses $\sim$ 95 distinct names with ≤ 4 occurrences per (name, emotion) cell; (ii) curate a 15-scenario pool per topic (120 total) and rotate as scenarios$[k \bmod 15]$, capping each scenario at ≤ 2 occurrences per cell; (iii) for WritingStyle, curate 15 subthemes per topic and rotate round-robin so each subtheme contributes 4 stimuli at full scale.

Number-stimulus design. The Number dataset extends the design in Hu et al. [2026] in two ways. First, prompts are *set-property* queries ("Are there any prime numbers?", "Are there any numbers greater than 5?") rather than value-identification queries, because lexical surface-form matching

trivially solves "Is the number exactly 3?" on the input "I have 3 apples" without engaging numerical processing. Second, the single-number random insertion is generalized to a balanced 1-/2-/3-number multi-number variant. Stimuli combine three types: (a) pseudo-stimuli (7-word `wikitext-103` chunks [Merity et al., 2016] with $c \in \{1, 2, 3\}$ target numbers inserted at random positions; N=90); (b) real Wikipedia sentences containing one or more target numbers 1–9 ($N \approx 675$); and (c) task-template stimuli reproducing five cognitive tasks:quantity, comparison, arithmetic, property, ordinal, with five phrasings per (task, number, format) cell ($N \approx 442$). Anchor values in comparison/arithmetic templates are mathematically validated against the target before acceptance.

Table 5: Example stimuli from **EmotionalStory** (topic $\times$ emotion, with one example per emotion and per topic, repeated to balance both axes; first $\sim$130 characters of each story shown). The full corpus is 8 emotions $\times$ 8 topics $\times$ 30 = 1,920 stories.

Emotion	Topic	Scenario	First $\sim$130 chars
joy	career	performance review meeting	"Aaliyah sat across from her manager, hands folded in her lap, bracing for the usual mix of praise and critique. Instead, her. . ."
sadness	education	a thesis defense	"Devon had spent five years on this research, and as the committee chair slid the marked-up pages across the table and quietly. . ."
anger	family	a kitchen scene	"Iris had asked her younger brother three times to stop leaving his dirty dishes in the sink, and yet there they were again — a. . ."
fear	friendship	a coffee shop meetup	"Nasir arrived at the coffee shop early, rehearsing what he needed to say to his best friend of ten years — that he had. . ."
trust	health	a doctor's visit	"Saif sat across from Dr. Okafor, who had been his physician for over a decade, and listened as she carefully explained his new. . ."
disgust	hobbies	a painting session	"Wilhelm set his brush down and leaned in to examine the still-life arrangement he'd been painting for the past hour — a bowl of. . ."
surprise	finance	buying a house	"Anika sat across from the loan officer, bracing herself for the familiar sting of rejection, when he slid a folder across the. . ."
anticipation	travel	an airport delay	"Felipe checked the departures board for the fourth time in ten minutes, watching the blinking status next to his flight number. . ."
joy	education	a thesis defense	"Beatriz stood at the front of the conference room, her hands trembling slightly as the last committee member lowered his pen and. . ."
sadness	family	a kitchen scene	"Gerwin stood at the kitchen counter, turning his mother's old recipe card over in his hands, the handwriting so familiar it made. . ."
anger	friendship	a coffee shop meetup	"Lila sat across from her best friend at their usual corner table, watching her scroll through her phone for the third time since. . ."
fear	health	a doctor's visit	"Ren sat on the crinkled paper of the examination table, hands gripping the edge until the knuckles went pale, waiting for the. . ."
trust	hobbies	a painting session	"Thandiwe set her half-finished canvas on the easel beside her mentor's, nervously watching as the older woman leaned in to study. . ."
disgust	finance	buying a house	"Adaeze flipped through the inspection report for the house she had been dreaming about for months, and her stomach turned when. . ."
surprise	travel	an airport delay	"Diego had settled into his gate seat with a coffee and a book, fully resigned to the four-hour delay the departures board had. . ."
anticipation	career	performance review meeting	"Magnus straightened his tie for the third time as he sat outside his manager's closed office door, the muffled voices inside. . ."

Prompt templates. EmotionalStory uses *"What emotion does this text express?"* and *"What topic is this text about?"* as the two open prompts; the sixteen specific prompts are yes/no queries *"Does*

Table 6: Example stimuli from **WritingStyle** (topic $\times$ style). The same neutral content seed is rendered in each of the six styles within a topic, so within-topic rows differ *only* in style. Each style and topic appears at least once; the full corpus is 6 styles $\times$ 4 topics $\times$ 60 $=$ 1,440 short passages.

Style	Topic	First sentence
formal	nature	"Forests occupy approximately 31 percent of Earth's total land surface and constitute the primary habitat for an estimated 80 percent of the planet's terrestrial biodiversity."
casual	nature	"Forests take up about 31% of all the land on Earth, which is pretty impressive when you think about it."
poetic	nature	"Like a living mantle draped across nearly a third of Earth's ancient skin, forests breathe and pulse with the world's wild abundance."
technical	technology	"A smartphone is a handheld electronic device that integrates mobile telephony with general-purpose computing capabilities, incorporating. . ."
journalistic	technology	"Smartphones are handheld devices that integrate mobile phone capabilities with computing functions, including internet access and. . ."
archaic	technology	"Behold the smartphone, that marvellous handheld contrivance which doth unite within its slender form the offices of the common. . ."
formal	food	"Macronutrients constitute a class of essential dietary compounds required by the human body in substantial quantities to support. . ."
casual	food	"Macronutrients are basically the nutrients your body needs a lot of to keep you energized and help you grow."
poetic	food	"Like pillars holding up the temple of the body, macronutrients are the great sustaining forces that fuel our every breath, our. . ."
technical	sports	"Each match consists of two opposing teams, each fielding exactly eleven players on the pitch simultaneously."
journalistic	sports	"Two teams of eleven players each compete in soccer, a sport in which the objective is to advance the ball into the opposing team's goal."
archaic	sports	"Verily, the noble sport of soccer is contested betwixt two sides, each comprising eleven players of valiant constitution."

Table 7: Example stimuli from **Number**, spanning the three stimulus types and the five task framings of Hu et al. [2026], each shown in both *digit* and *word* numerical formats. The full corpus has $N \approx 1{,}247$ stimuli.

Stimulus type	Task framing	Format	Text
pseudo-stimulus	—	digit	"Sold at a lower price than 7 the."
pseudo-stimulus	—	word	"Museum remained eight in the tower building for."
real Wikipedia	—	digit	"An academic study found that Jordan's 1st NBA comeback resulted in an increase in the market. . ."
real Wikipedia	—	word	"In 2014, Fernandez played the leading lady in Sajid Nadiadwala's *Kick*, which is one of the. . ."
task-template	quantity	digit	"I have a total of 1 apple."
task-template	quantity	word	"I have a total of one apple."
task-template	comparison	digit	"A number smaller than 15 is 1."
task-template	comparison	word	"A number smaller than nine is one."
task-template	arithmetic	digit	"The sum of 0 and 1 is 1."
task-template	arithmetic	word	"The sum of zero and one is one."
task-template	property	digit	"A perfect square is 1."
task-template	property	word	"An example of an odd number is one."
task-template	ordinal	digit	"1 comes after 0."
task-template	ordinal	word	"Before two comes one."

this text express {value}?" (joy, sadness, anger, fear, trust, disgust, surprise, anticipation) and *"Is this about {value}?"* (career, education, family, friendship, health, hobbies, finance, travel). WritingStyle uses the analogous open pair *"What writing style is this in?"* / *"What topic is this text about?"* with the six styles ({formal, casual, poetic, technical, journalistic, archaic}) and four topics ({nature, technology, food, sports}). Number uses the nine set-property prompts distributed across the five cognitive tasks (1 quantity / 3 comparison / 2 arithmetic / 2 property / 1 ordinal). Irrelevant prompts (*"What is the capital of France?"* and similar) are shared across all datasets. **StyleTransfer** uses *"What scene does the image depict?"* and *"What artistic style does the image belong to?"* as the open pair, with eleven specific prompts: four scenes (*"Is this a {beach, bedroom, library, mountain}?"*) and seven styles (*"Is this in the style of {Demuth, Klimt, Monet, Pollock}?"*). *"Are there people*

in this image?" (detection) and *"How many people are in this image?"* (count) as the open pair, with eight count-specific yes/no queries of the form *"Are there exactly {0, 1, 2, 3, 4, 5} people in this image?"* plus the two range bins *"Are there between 5 and 10 people?"* and *"Are there more than 10 people?"*. **EmoSet** (VLM) uses *"What is depicted in this image?"* (content) and *"What emotion does this image evoke?"* (emotion) as the open pair, with twelve specific prompts: four content categories (*"Are there people / animals?"*, *"Is this a natural landscape?"*, *"Is this primarily an inanimate scene?"*) and eight emotions (*"Does this image evoke {amusement, awe, contentment, excitement, anger, disgust, fear, sadness}?"*). Irrelevant prompts (*"What is the capital of France?"*, *"What is 2+2?"*, *"What day of the week is it?"*, and seven similar factual-knowledge questions) are shared across all six datasets. Full prompt JSONs will be in the released code.

F Implementation details

Layer indexing. For each model we extract hidden states at every transformer block at the last input-prompt token. Layer indices are reported on a normalized $[0, 1]$ scale so that models of different depth can be overlaid; absolute layer counts are: OPT-2.7B (32), Llama-3-8B-Instruct (32), Qwen3-8B (36), BLIP-2 (32), LLaVA-OneVision-7B (28), Qwen3-VL-8B (36).

Feature extraction. For each (model, dataset, prompt) cell we feed every stimulus through the model under the prompt and store the residual-stream hidden state at the *last input-prompt token* from every transformer block; this token is the position at which the model commits to a continuation and is the conventional probe site for prompt-conditioned reads. LLM inputs are formatted using the model's training-time convention: chat models (Llama-3-Instruct, Qwen-2-Instruct, Qwen3) use `tokenizer.apply_chat_template(..., add_generation_prompt=True)`; OPT-2.7B (a non-instruction-tuned base model) uses the QA template (`Question: {prompt}\nText: {text}\nAnswer:`). VLM inputs are processed by each model's native processor (BLIP-2, LLaVA-OneVision, Qwen3-VL). Hidden states are stored as zarr arrays of shape $(1,\, n_{\text{stim}},\, D)$ per layer with metadata recording the pooling convention, model identity, prompt string, sequence length, and the number of truncated stimuli (typically zero).

Cross-validation, ridge, and MLP. Five-fold stratified cross-validation across stimuli, ridge $\lambda{=}1$, and a one-hidden-layer MLP ($H{=}512$, GELU, AdamW with weight decay 10^{-3}, 200 epochs at mini-batch size ≤ 256), see Appendix A for the estimator definitions.

Behavioural evaluation. Per-dataset keyword evaluators check whether generated text satisfies the target prompt's semantic constraint, scoring two metrics: a *relevance* flag (does the text address the target attribute at all, e.g. does it mention any artistic-style or emotion vocabulary) and an *accuracy* flag (does the text identify the correct ground-truth label, e.g. "Monet" for a Monet-style image, "joy" for a joy-emotion story, the correct count for COCO). For style and emotion prompts the evaluator combines artist-name keywords, art-school keywords, and synonym lists (e.g. "post-impressionism", "van gogh", "starry night" all map to `vangogh`). Evaluator dictionaries will be in the released code.

G Audit of the behavioural scorer

Keyword matching is not self-evidently sufficient for scoring open-ended generations, so we audited the per-dataset keyword evaluators of Appendix F against two independent LLM judges. The audit covered all 24,582 end-to-end prompt-B responses collected over every pair of model and dataset. Each response was scored three times: once by the original keyword dictionary, once by GPT-4o-mini, and once by the stronger judge GPT-5.6, which we take as the reference in Table 8.

Table 8: Agreement with the stronger judge (GPT-5.6) over all 24,582 end-to-end prompt-B responses.

Scorer	Agreement with GPT-5.6
Original keyword dictionary	82.3%
GPT-4o-mini judge	78.9%
Revised keyword dictionary	91.6%

Two observations follow. First, the original dictionary agreed with the stronger judge more often than the smaller LLM judge did, so replacing the dictionary with a small LLM judge would not have improved agreement on this corpus. Second, manual inspection of the disagreements traced them to three systematic dictionary rules: substring matching, literal treatment of binned labels, and mishandled multilabel responses. The correction consists of rewriting exactly those three rules, which raises agreement with the stronger judge from 82.3% to 91.6%.

Residual disagreement is not one-sided and remains comparable in volume in both directions: 963 responses are scored correct by the dictionary and wrong by GPT-5.6, against 1,102 responses in the opposite direction. We retain a dictionary scorer because it is deterministic, fully inspectable, and inexpensive to reproduce, and all behavioural numbers reported here use the revised version. The audit nevertheless places one residual limitation on record: a response that paraphrases stimulus content, or that continues a dialogue of its own, can still receive credit.

H Sample size and ridge regularization

The nominal parameter count of the more expressive tiers is large relative to the number of stimuli, because the representation width of these models is of the same order as the stimulus count. The first safeguard is built into the score itself: every reported R^2 is evaluated on held-out stimuli under five-fold cross-validation, so a map that merely reproduces the training stimuli cannot inflate the reported value. Overfitting would surface as a reduced or negative held-out R^2 rather than as a spuriously high one.

We also ran the direct control. For two pairs of model and dataset we regenerated the stimulus set at a substantially larger size and refit every tier under the identical protocol, as summarized in Table 9.

Table 9: Enlarged stimulus sets used for the sample-size control, with the representation width of each model for comparison.

Model	Dataset	Submitted	Enlarged	Width
Llama-3-8B-Instruct	EmotionalStory	1920	5760	4096
LLaVA-OneVision	COCO	1000	5000	3584

Across these enlarged sets the cumulative held-out R^2 over the deepest half of the layers changes by at most 0.023, and the per-prompt-pair medians move upward rather than downward. The direction matters as much as the magnitude: a fit sustained by a stimulus count too small for the hidden width would be expected to deteriorate once the sample is enlarged, whereas here it improves slightly. Finally, we replaced the fixed ridge penalty of Appendix F with nested cross-validation, selecting the penalty on an inner split of the training folds only. This leaves the affine increment essentially unchanged, so the affine tier's advantage is not an artifact of the particular penalty value used in the main analysis.

I Normalized prompt-induced activation changes

The normalized squared Frobenius distance $\|\mathbf{X}^B - \mathbf{X}^A\|_F^2 / \|\mathbf{X}^A\|_F^2$ are shown separately for the six prompt-pair groups (G1–G6). The summary across models and datasets is shown in Fig. 7, and the individual (model, dataset) pair results are in Figs. 8–13. The two irrelevant-source prompt groups (G5–G6) produce the largest activation change, followed by the within-attribute open→specific pair (G3) and the two cross-attribute pairs (G1: open→open; G2: specific→specific). The within-attribute specific→specific pair (G4) produces the smallest change, consistent with both prompts querying the same attribute axis at the same specificity level.

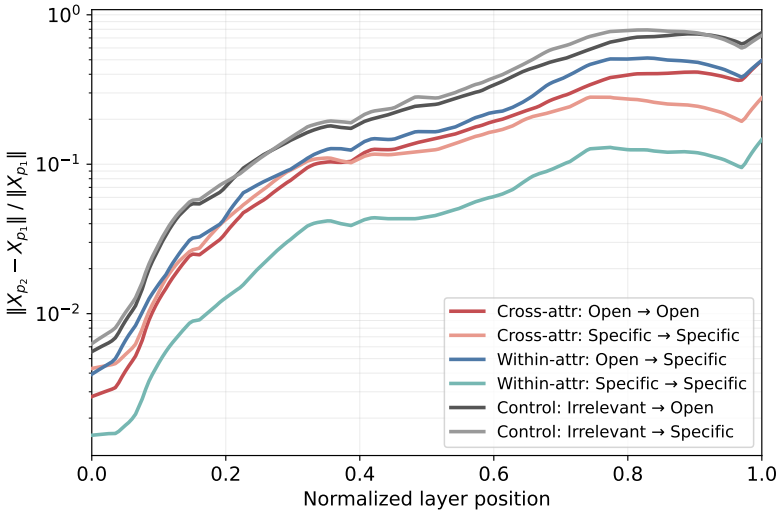

Figure 7: Normalized prompt-induced activation change across the six prompt-pair groups, pooled across all (model, dataset) cells. The depth-graded growth of the change is consistent across groups.

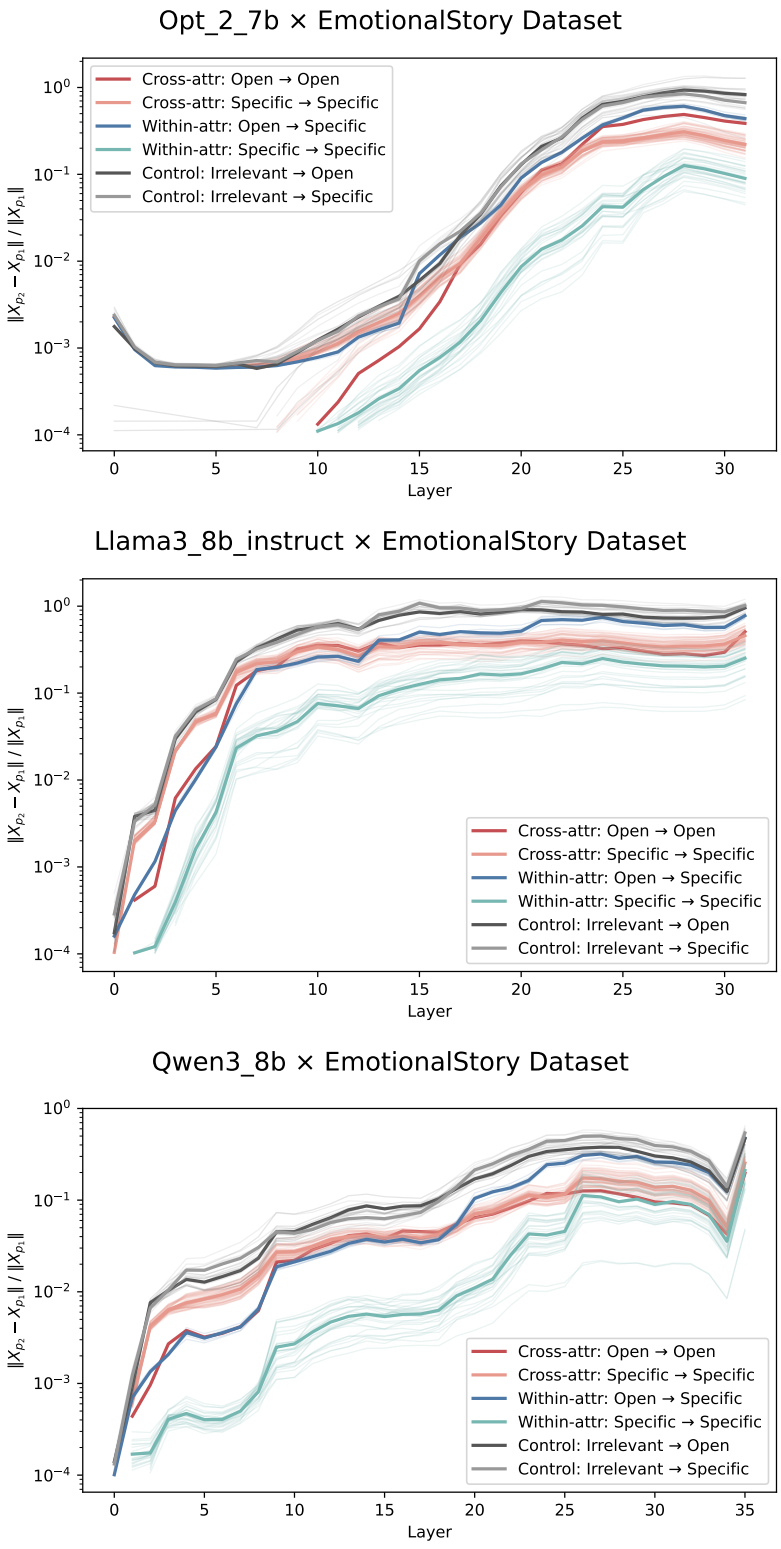

Figure 8: Normalized activation change for EmotionalStory across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

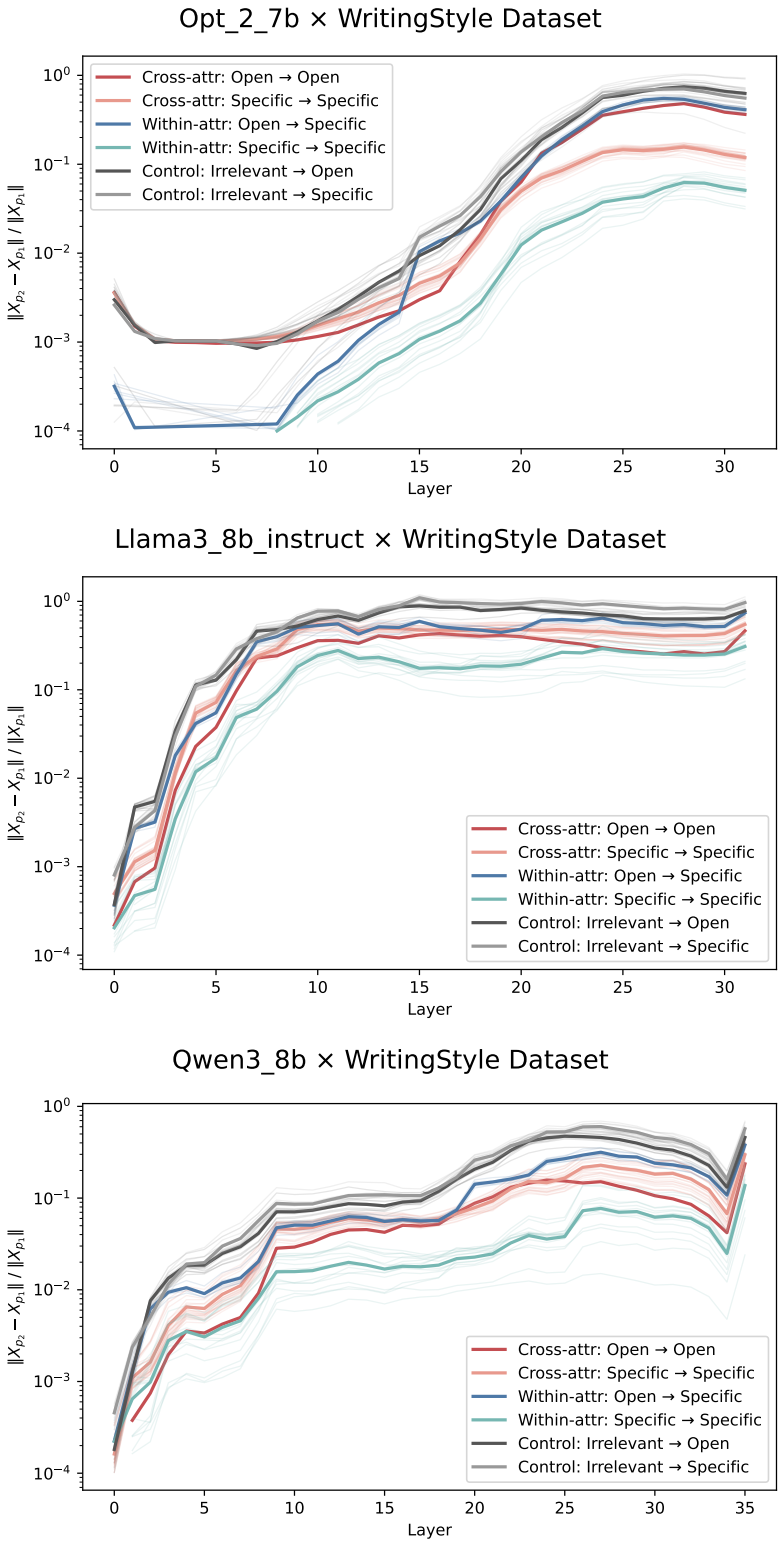

Figure 9: Normalized activation change for WritingStyle across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

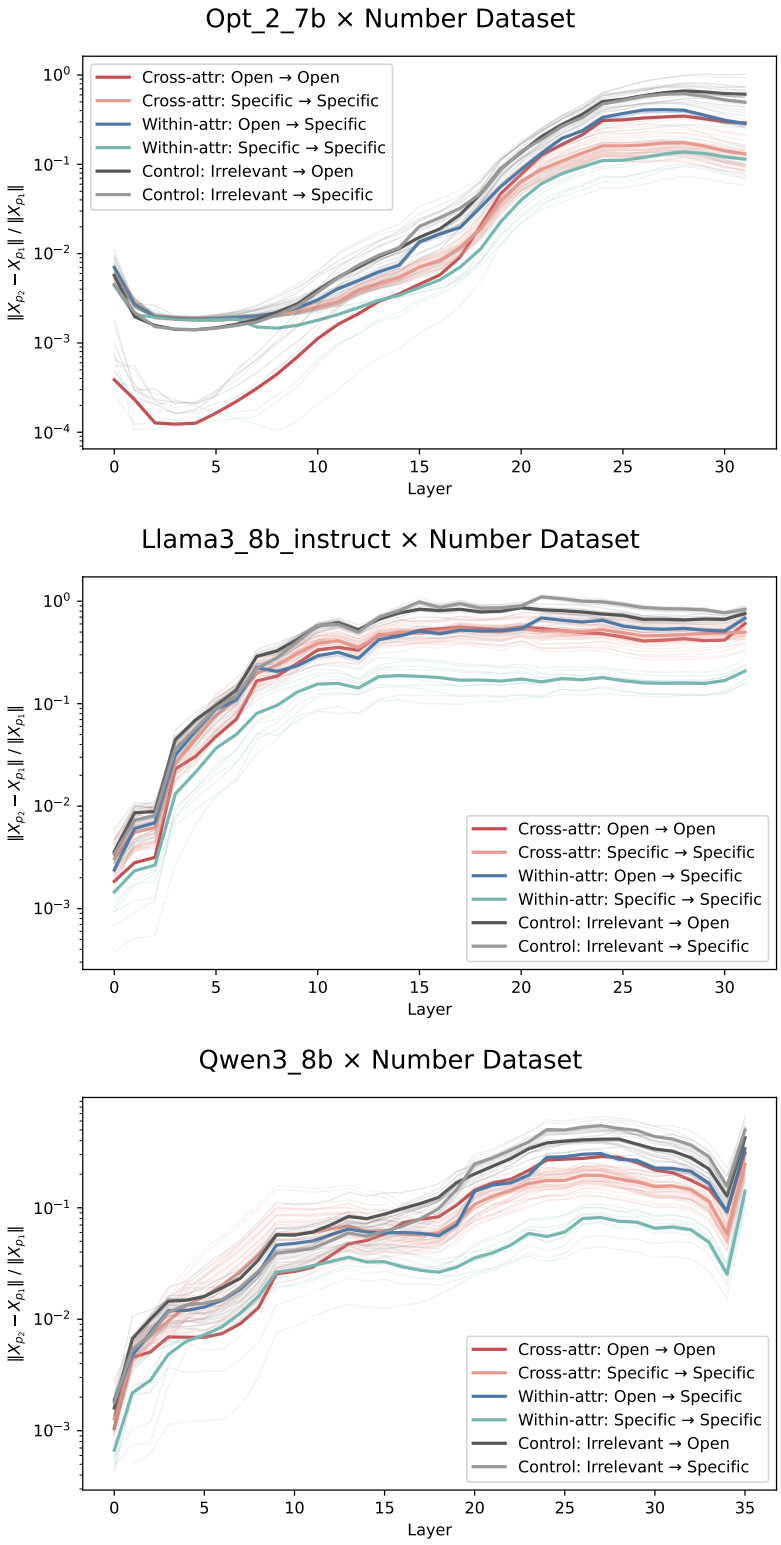

Figure 10: Normalized activation change for Number across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

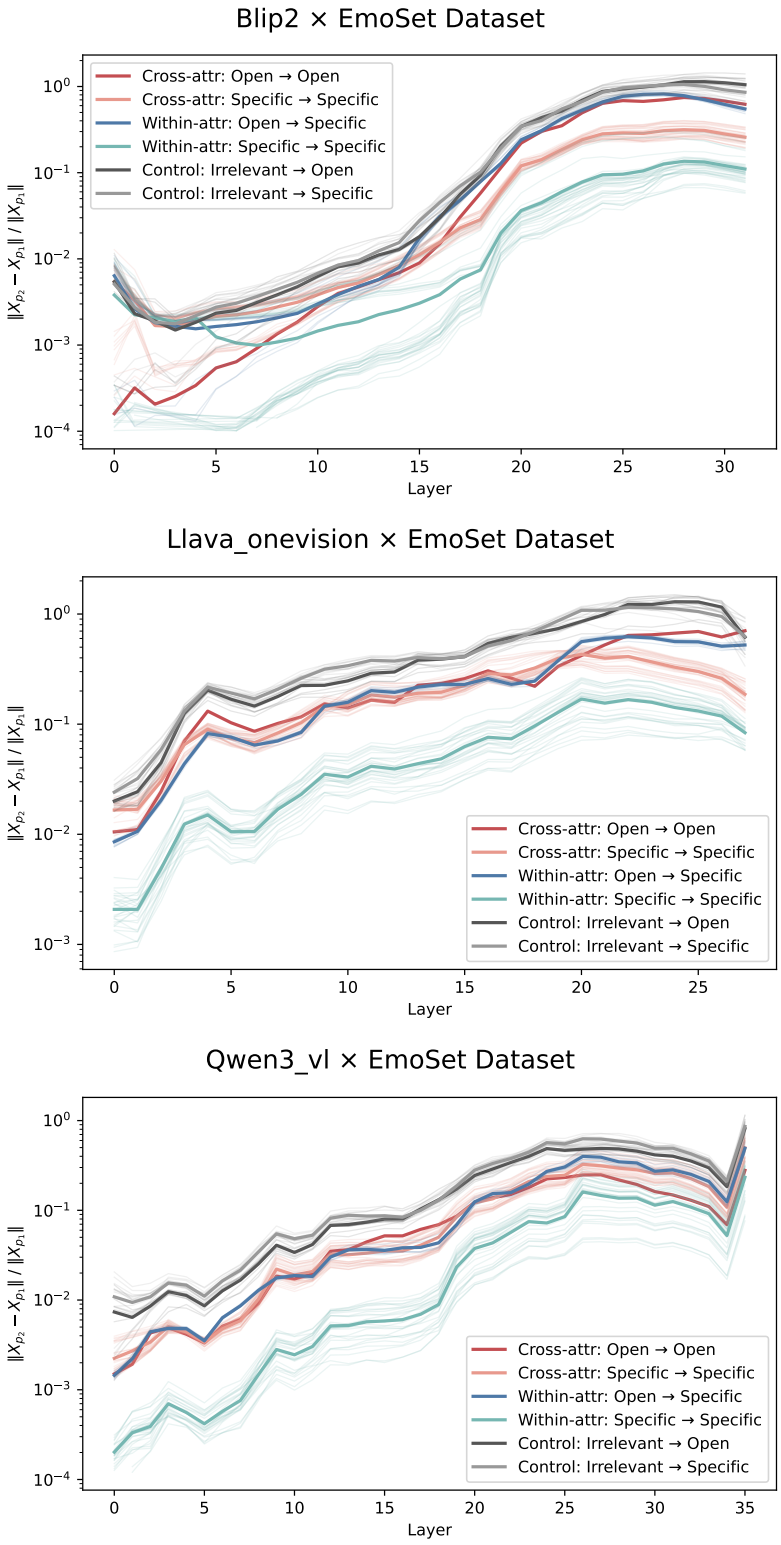

Figure 11: Normalized activation change for EmoSet across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).

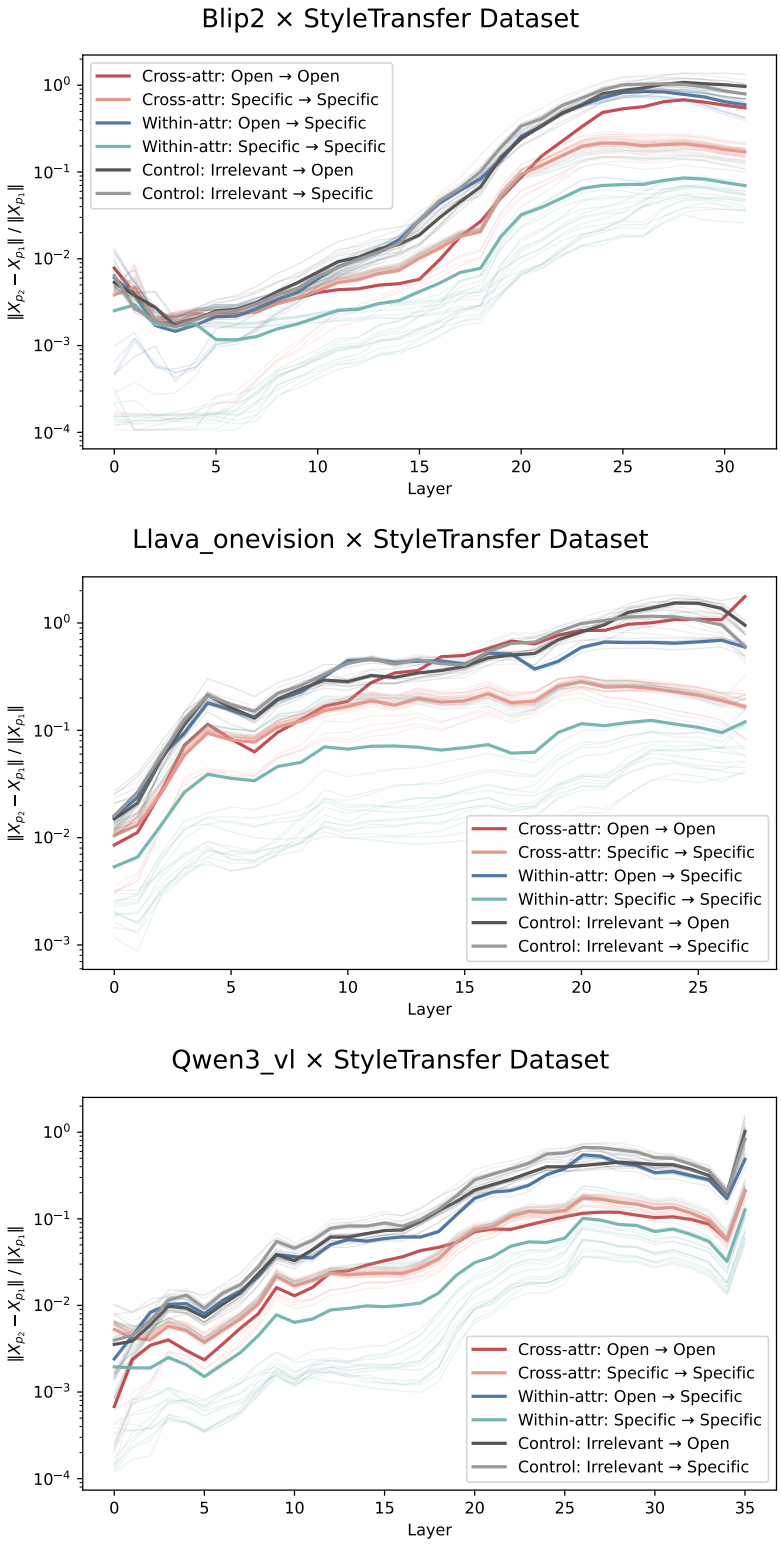

Figure 12: Normalised activation change for StyleTransfer across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).

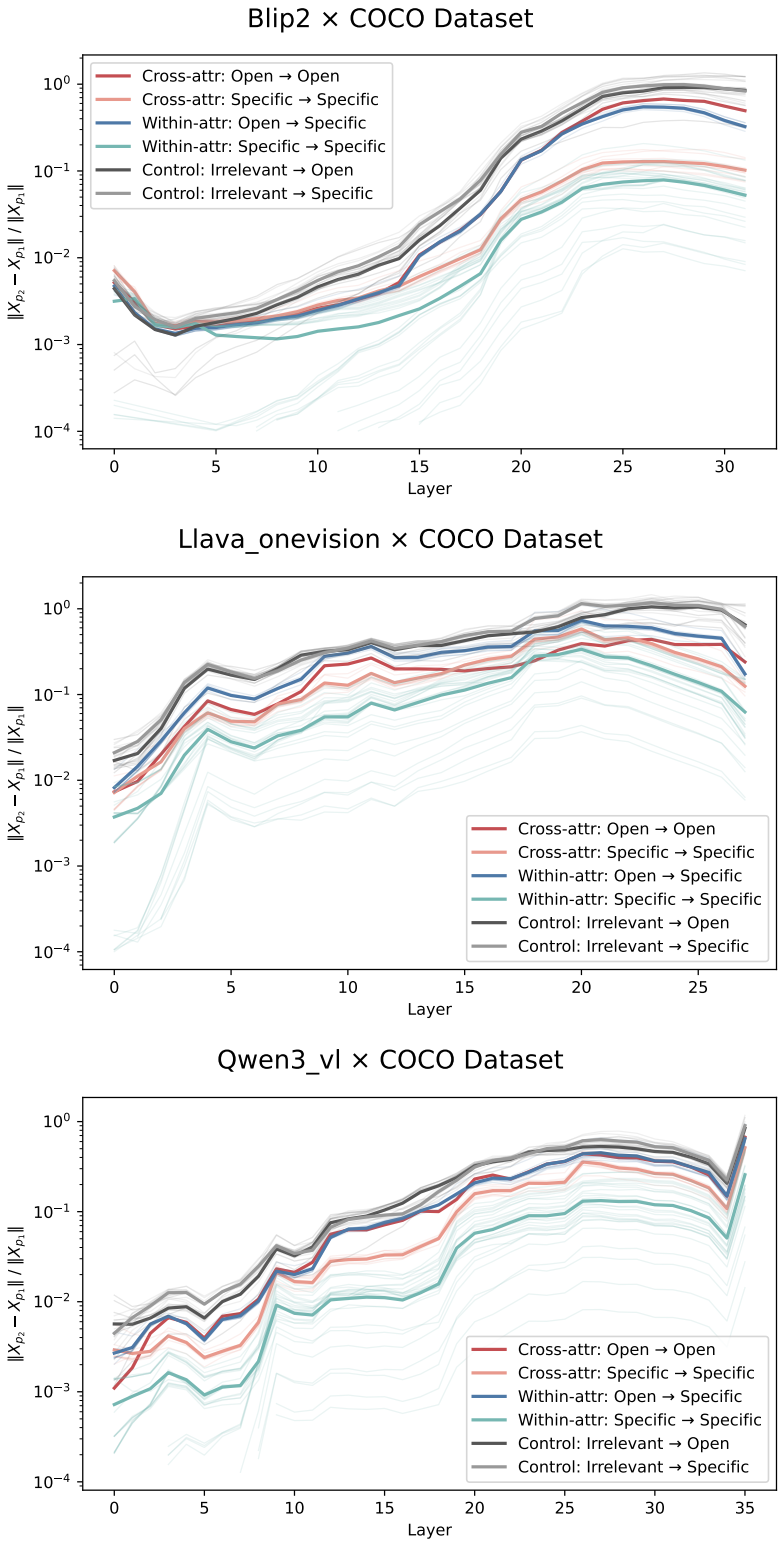

Figure 13: Normalized activation change for COCO across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).

J MDS visualizations

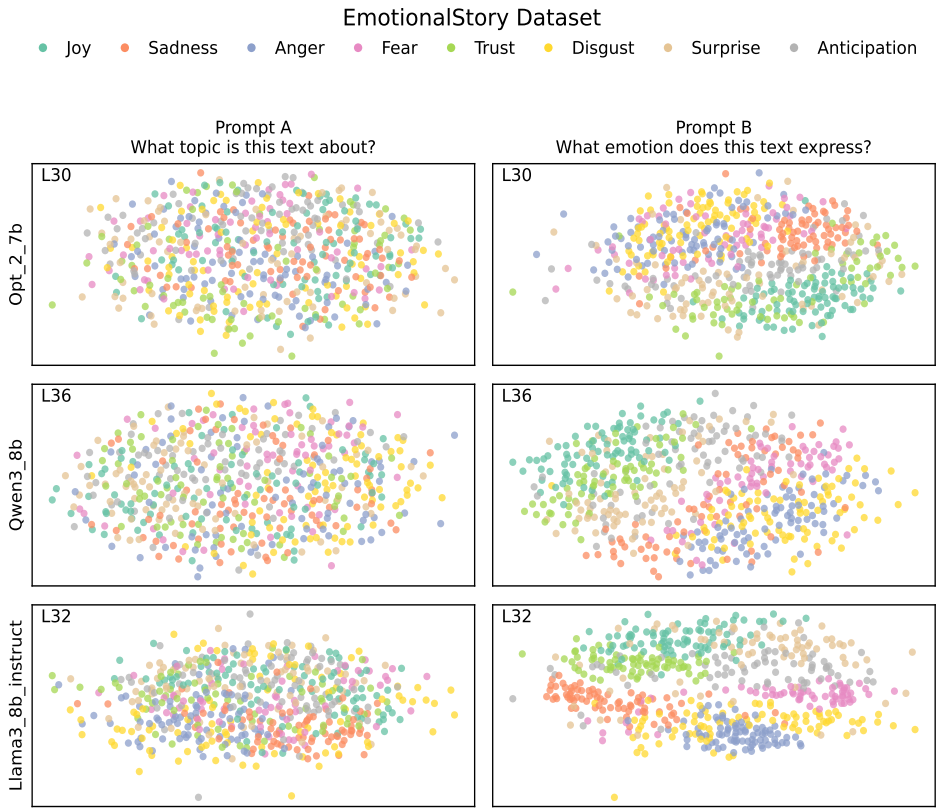

Figure 14: EmotionalStory dataset, prompt A (topic) vs. prompt B (emotion); stimuli coloured by ground-truth emotion.



Figure 15: WritingStyle dataset, prompt A (topic) vs. prompt B (writing style); stimuli coloured by ground-truth style.



Figure 16: Number dataset, prompt A (numbers mentioned) vs. prompt B (cognitive operation); stimuli coloured by ground-truth task framing.



Figure 17: EmoSet dataset, prompt A (image content) vs. prompt B (emotion evoked); stimuli coloured by ground-truth emotion.

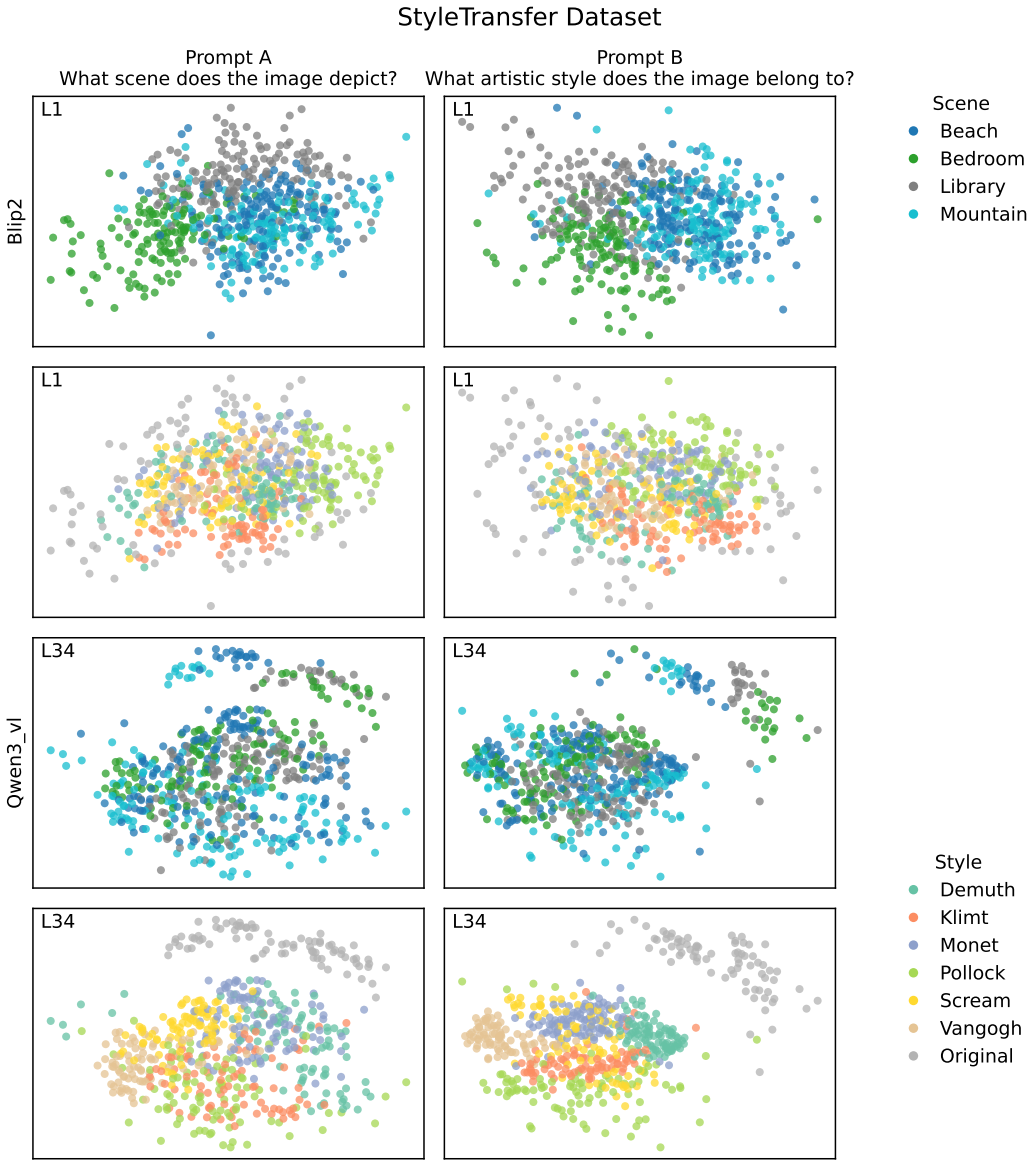

Figure 18: StyleTransfer dataset, prompt A (scene) vs. prompt B (artistic style); stimuli coloured by ground-truth scene (top legend) and style (bottom legend).

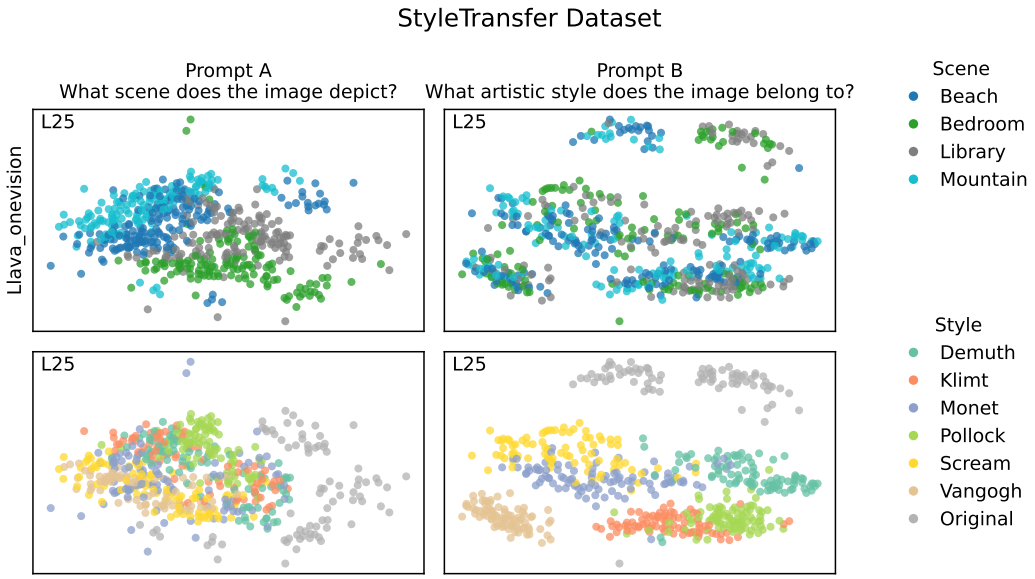

Figure 19: StyleTransfer dataset (continued), prompt A (scene) vs. prompt B (artistic style); stimuli coloured by ground-truth scene (top) and style (bottom).

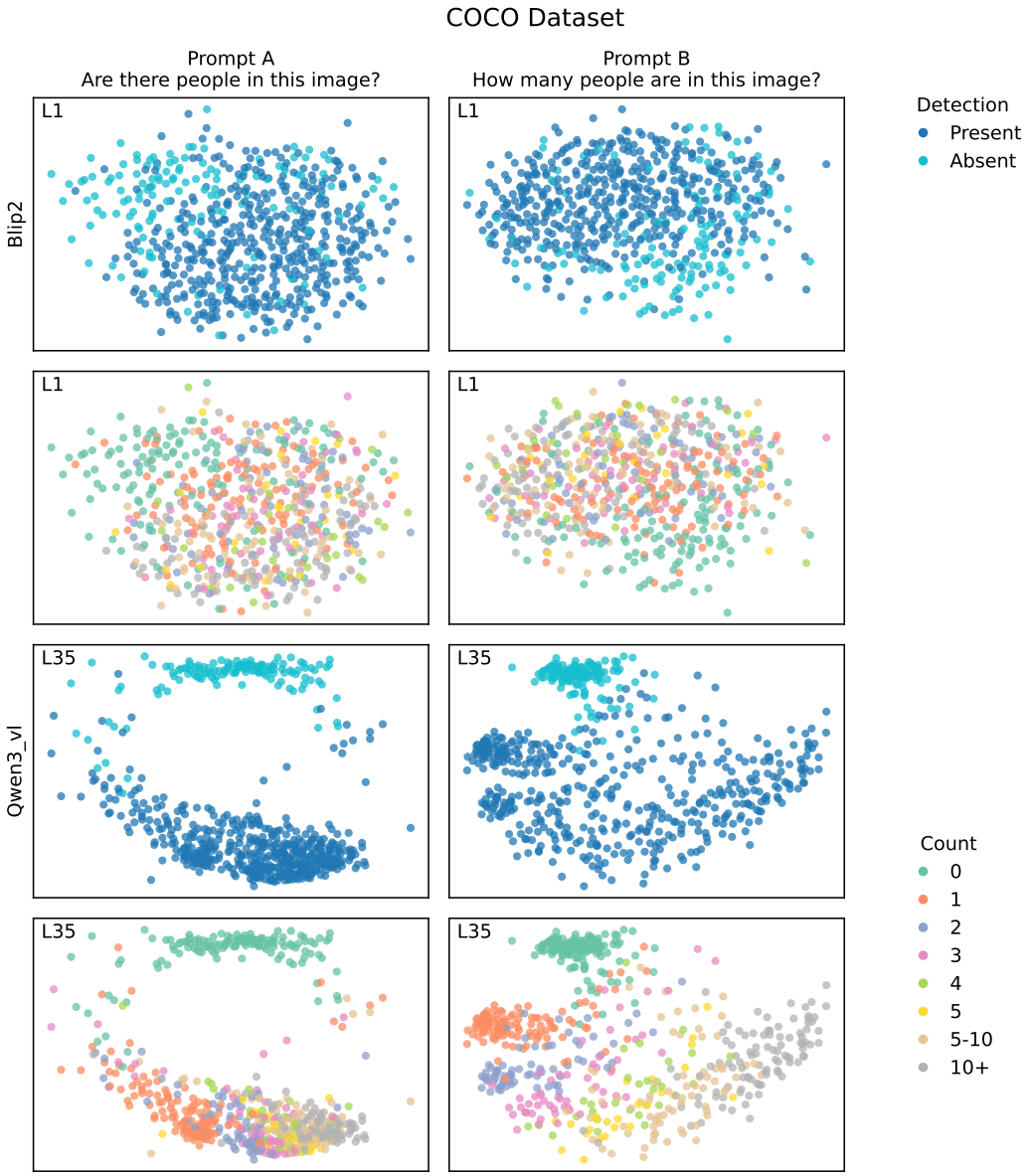

Figure 20: COCO dataset, prompt A (people detection) vs. prompt B (people count); stimuli coloured by ground-truth detection (top) and count bin (bottom).

K RSA and silhouette score

For each pair of model and dataset, we report the layerwise Spearman correlation between the data RDM of the prompt-induced hidden states and the prompt-B target attribute RDM, alongside the silhouette score based on the target-labels.

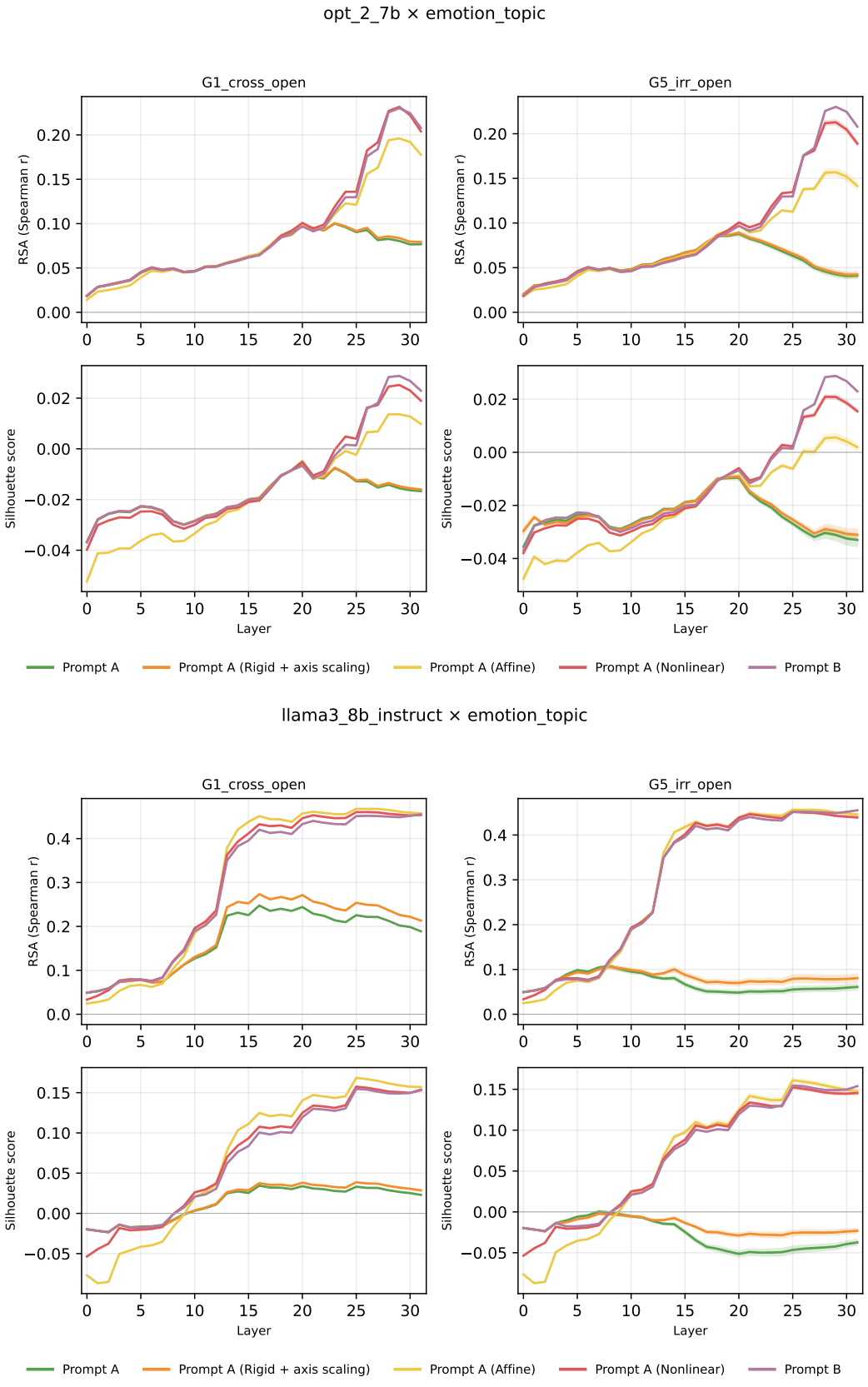

Figure 21: Layerwise RDM correlation (left) and silhouette score (right) for EmotionalStory (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

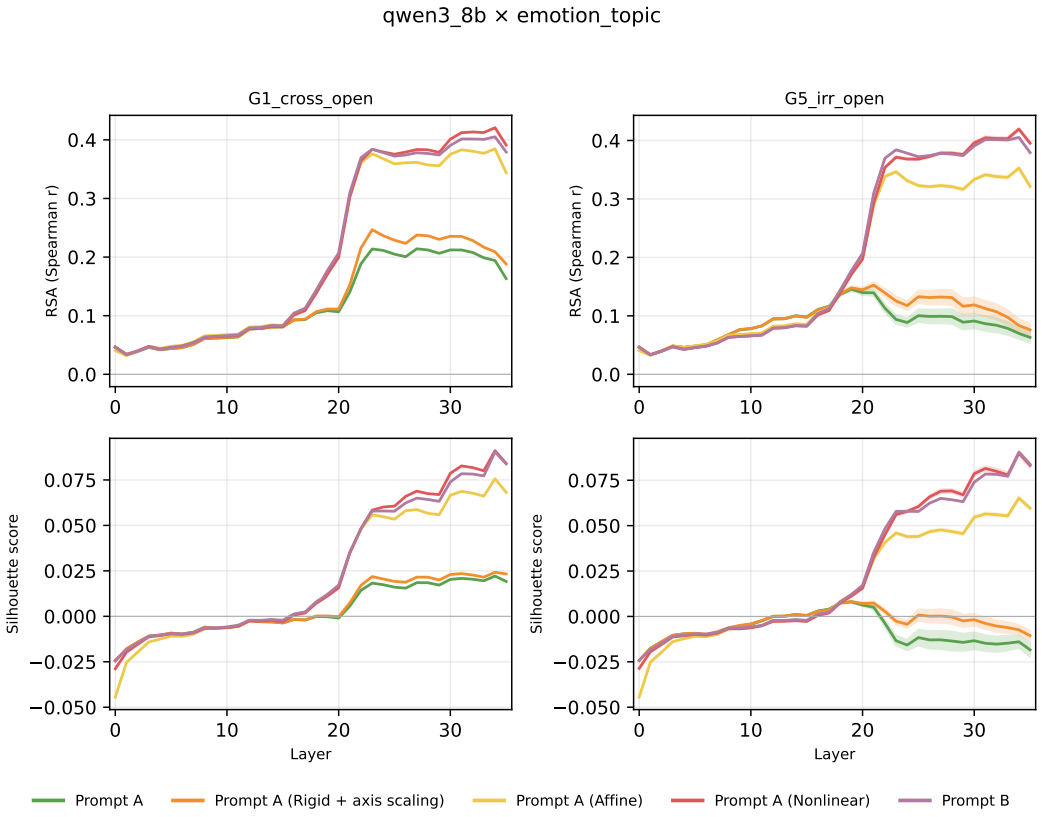

Figure 22: Layerwise RDM correlation (left) and silhouette score (right) for EmotionalStory (2/2) under prompt A vs. prompt B for for Qwen3-8B.

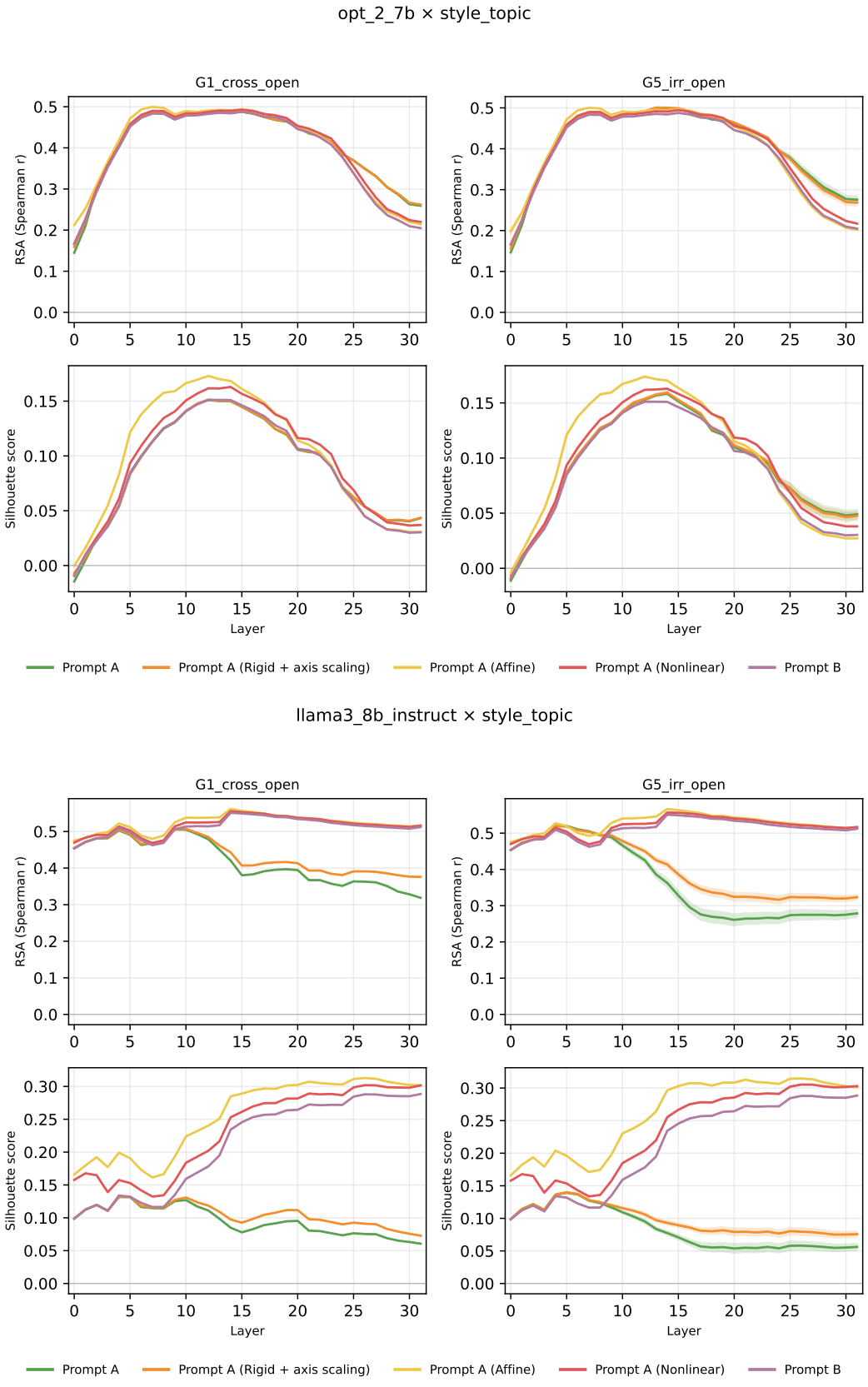

Figure 23: Layerwise RDM correlation (left) and silhouette score (right) for WritingStyle (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

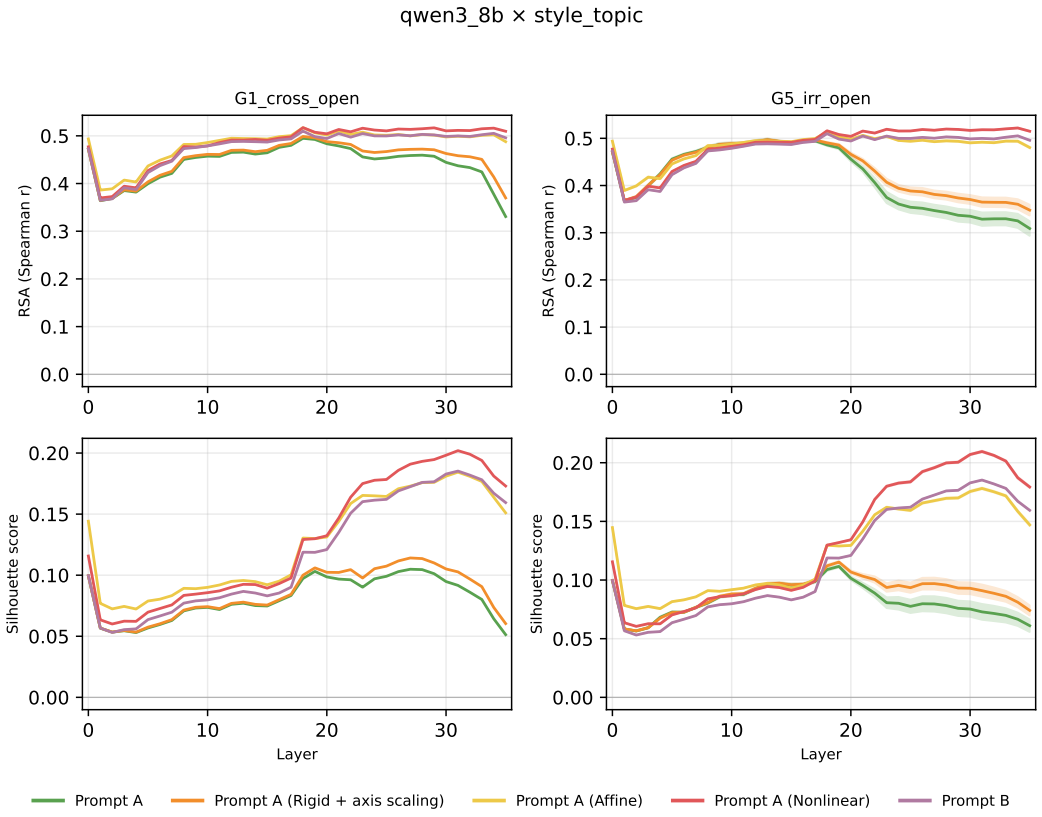

Figure 24: Layerwise RDM correlation (left) and silhouette score (right) for WritingStyle (2/2) under prompt A vs. prompt B for for Qwen3-8B.

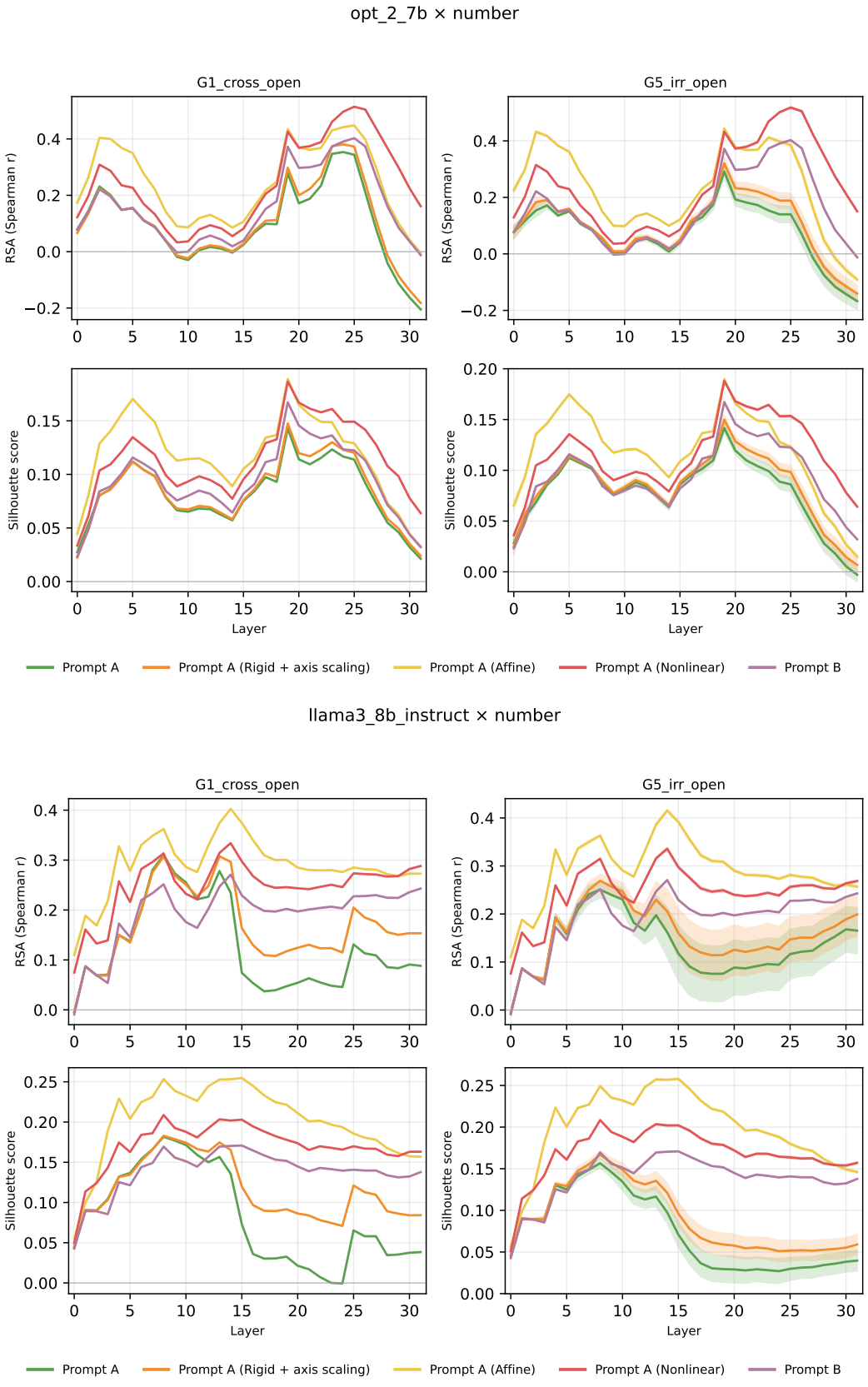

Figure 25: Layerwise RDM correlation (left) and silhouette score (right) for Number (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

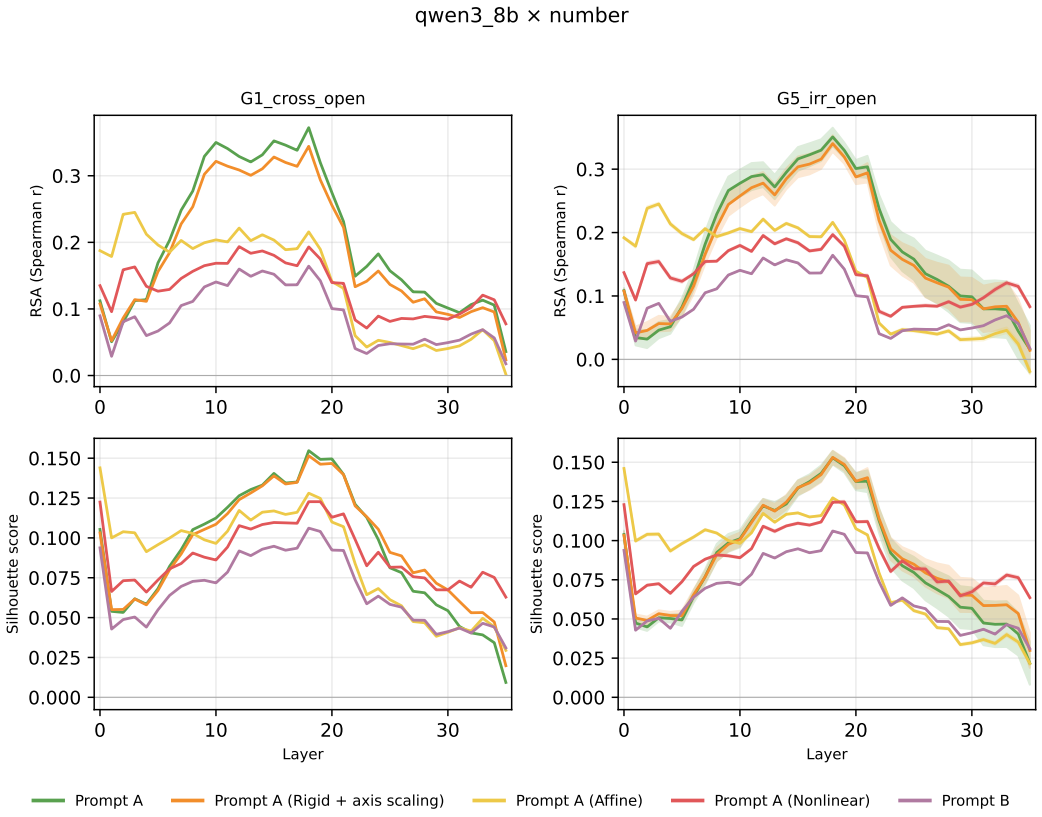

Figure 26: Layerwise RDM correlation (left) and silhouette score (right) for Number (2/2) under prompt A vs. prompt B for for Qwen3-8B.

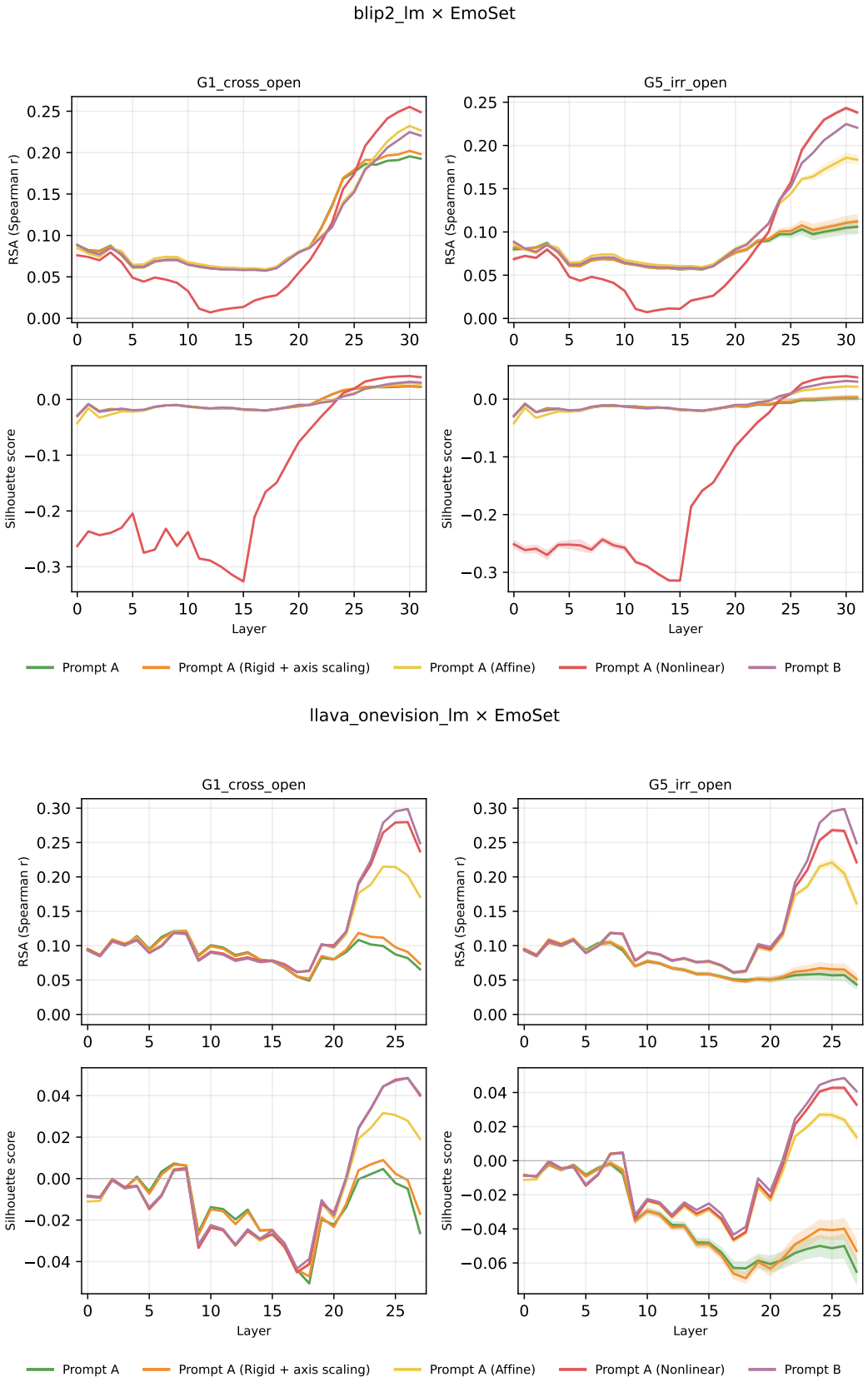

Figure 27: Layerwise RDM correlation (left) and silhouette score (right) for EmoSet (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

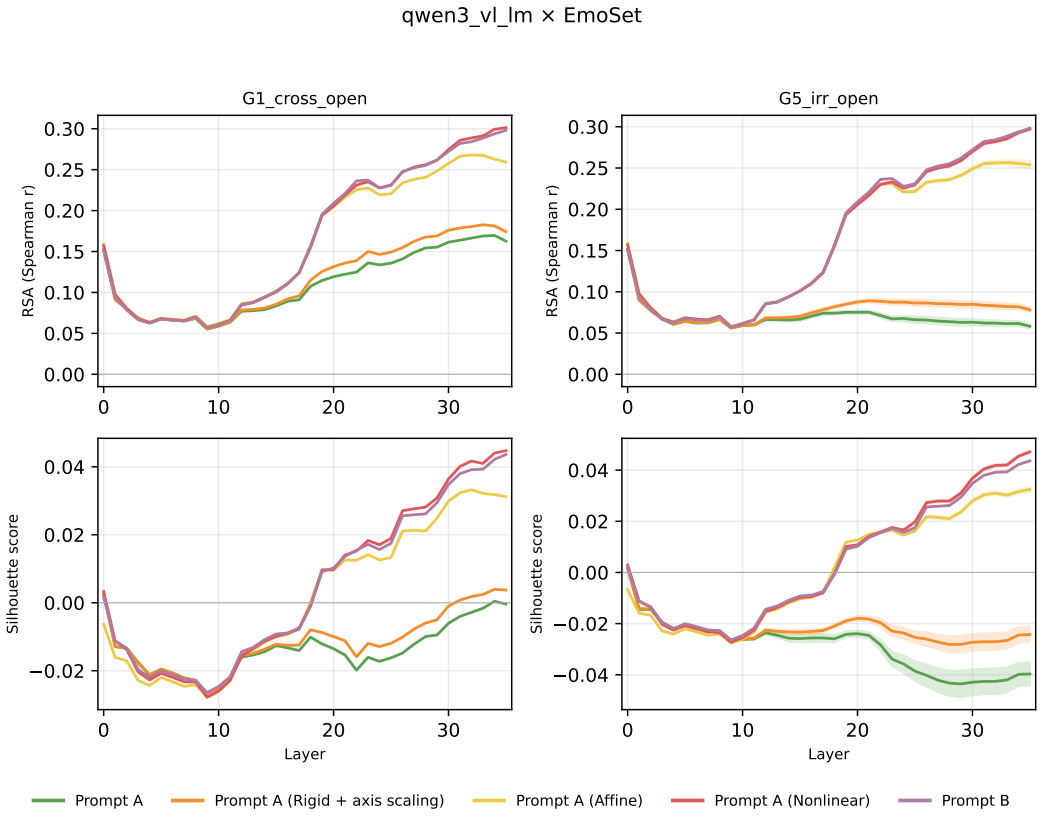

Figure 28: Layerwise RDM correlation (left) and silhouette score (right) for EmoSet (2/2) under prompt A vs. prompt B for for Qwen3-8B.

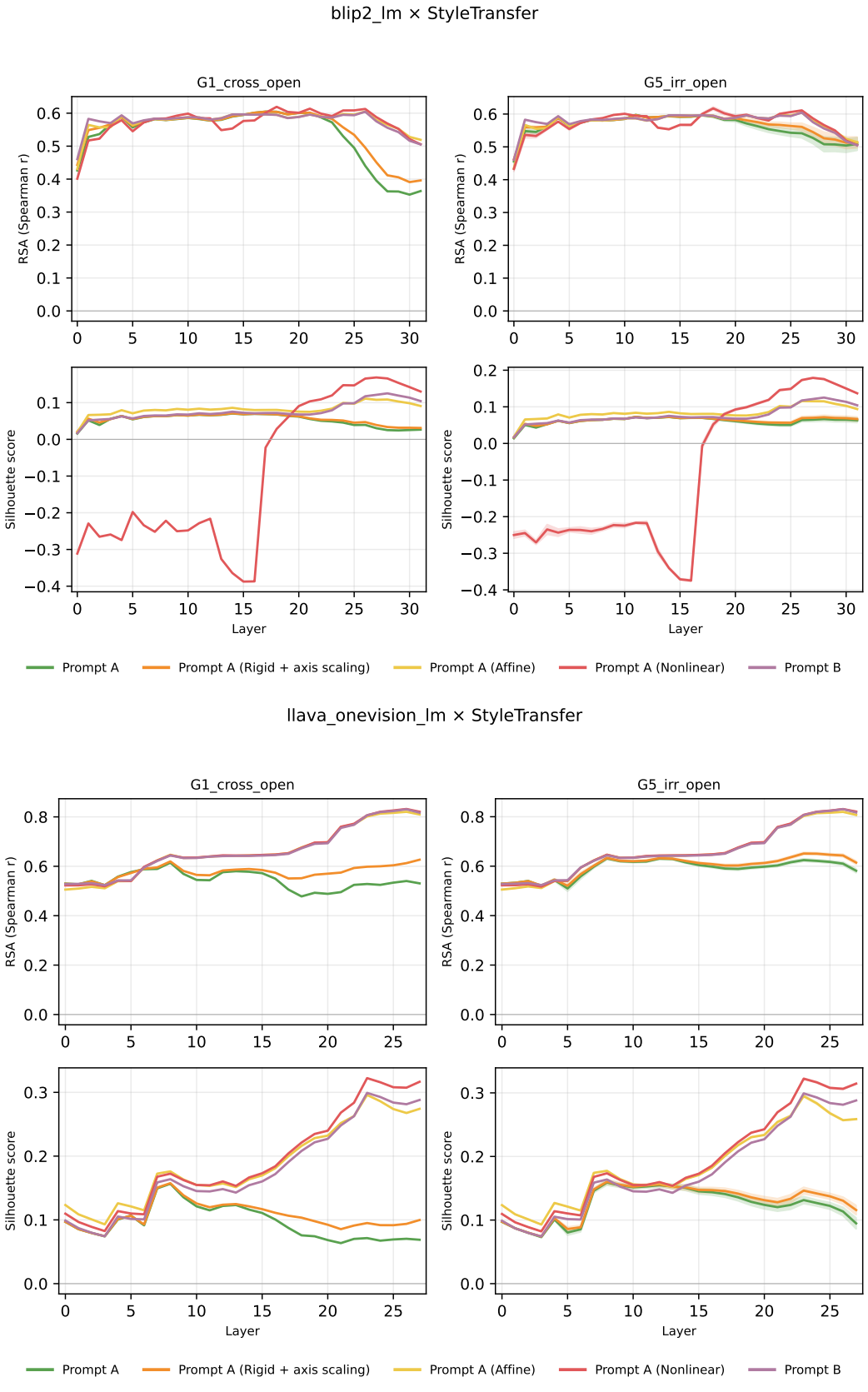

Figure 29: Layerwise RDM correlation (left) and silhouette score (right) for StyleTransfer (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

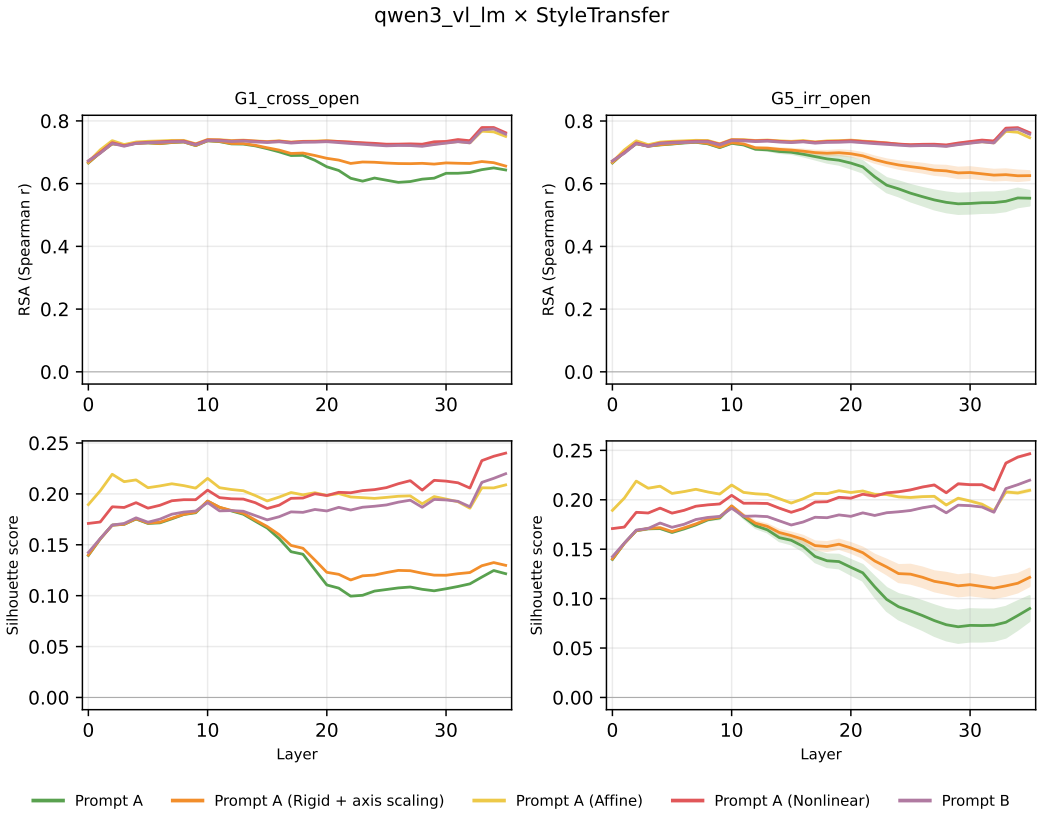

Figure 30: Layerwise RDM correlation (left) and silhouette score (right) for StyleTransfer (2/2) under prompt A vs. prompt B for for Qwen3-8B.

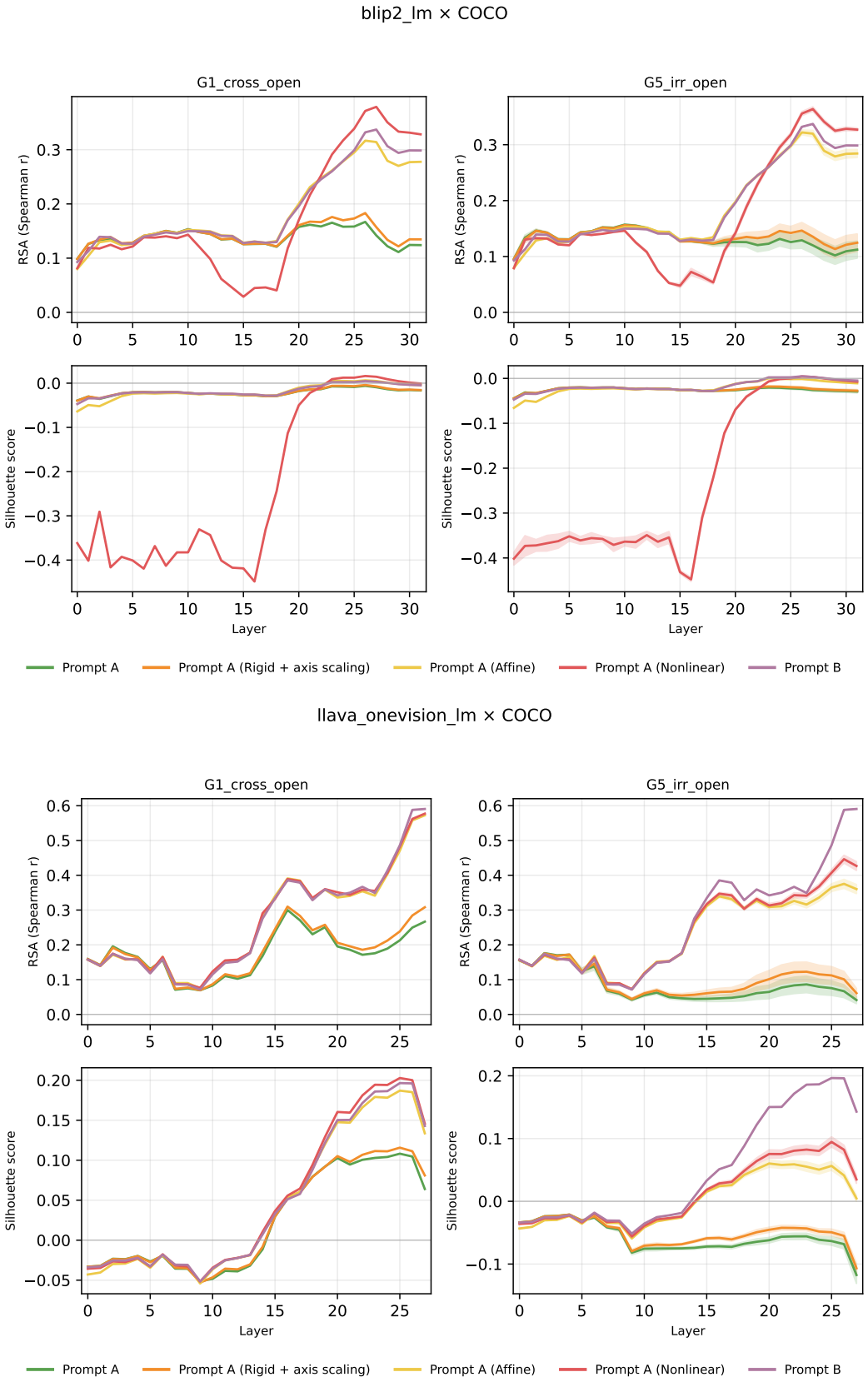

Figure 31: Layerwise RDM correlation (left) and silhouette score (right) for COCO (1/2) under prompt A vs. prompt B for OPT-2.7B (top) and Llama-3-8B-Instruct (bottom).

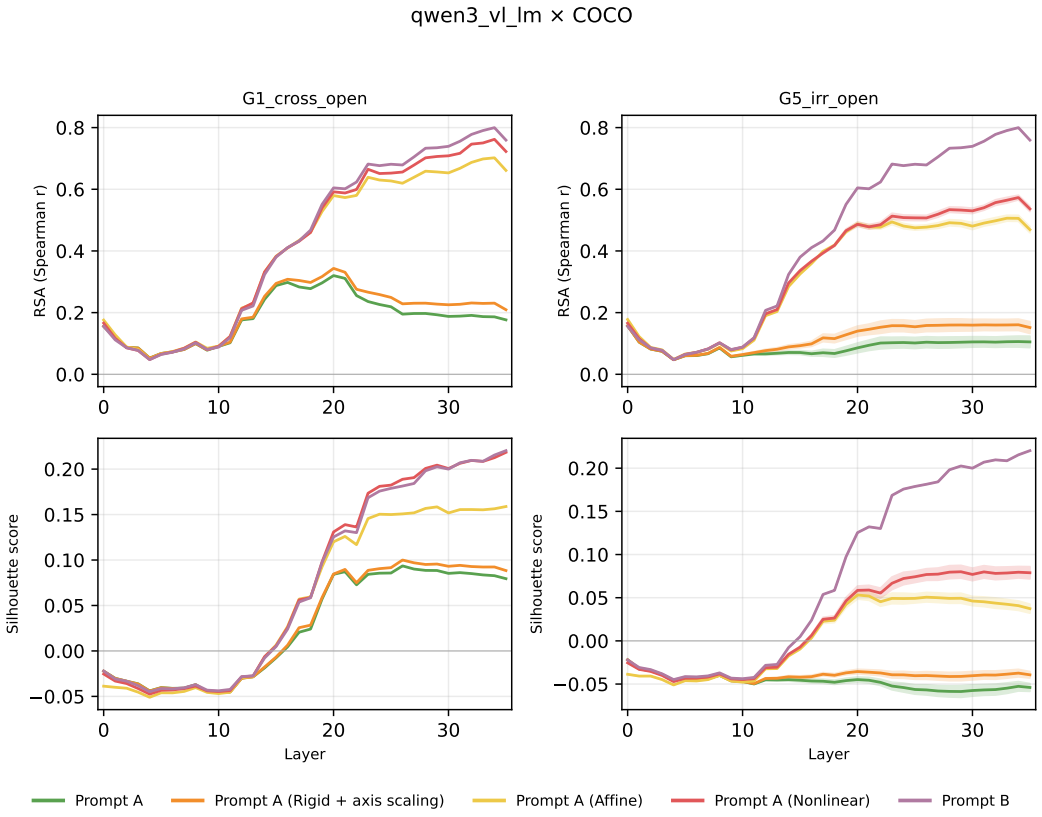

Figure 32: Layerwise RDM correlation (left) and silhouette score (right) for COCO (2/2) under prompt A vs. prompt B for for Qwen3-8B.

L Decomposition results for individual prompt-pair groups

For each (model, dataset) cell we show the full five-tier incremental cross-validated R^2 at every layer as a stacked bar. Each segment is $\Delta R^2_k = R^2_k - R^2_{k-1}$ for one tier in the nested chain (translation T, rigid transformation with uniform scaling O_u, rigid transformation with axis-wise scaling O_a, affine L, nonlinear N). These visualizations complement the summary Fig. 3 and make visible the *strategy-level* differences across model families discussed in §6.

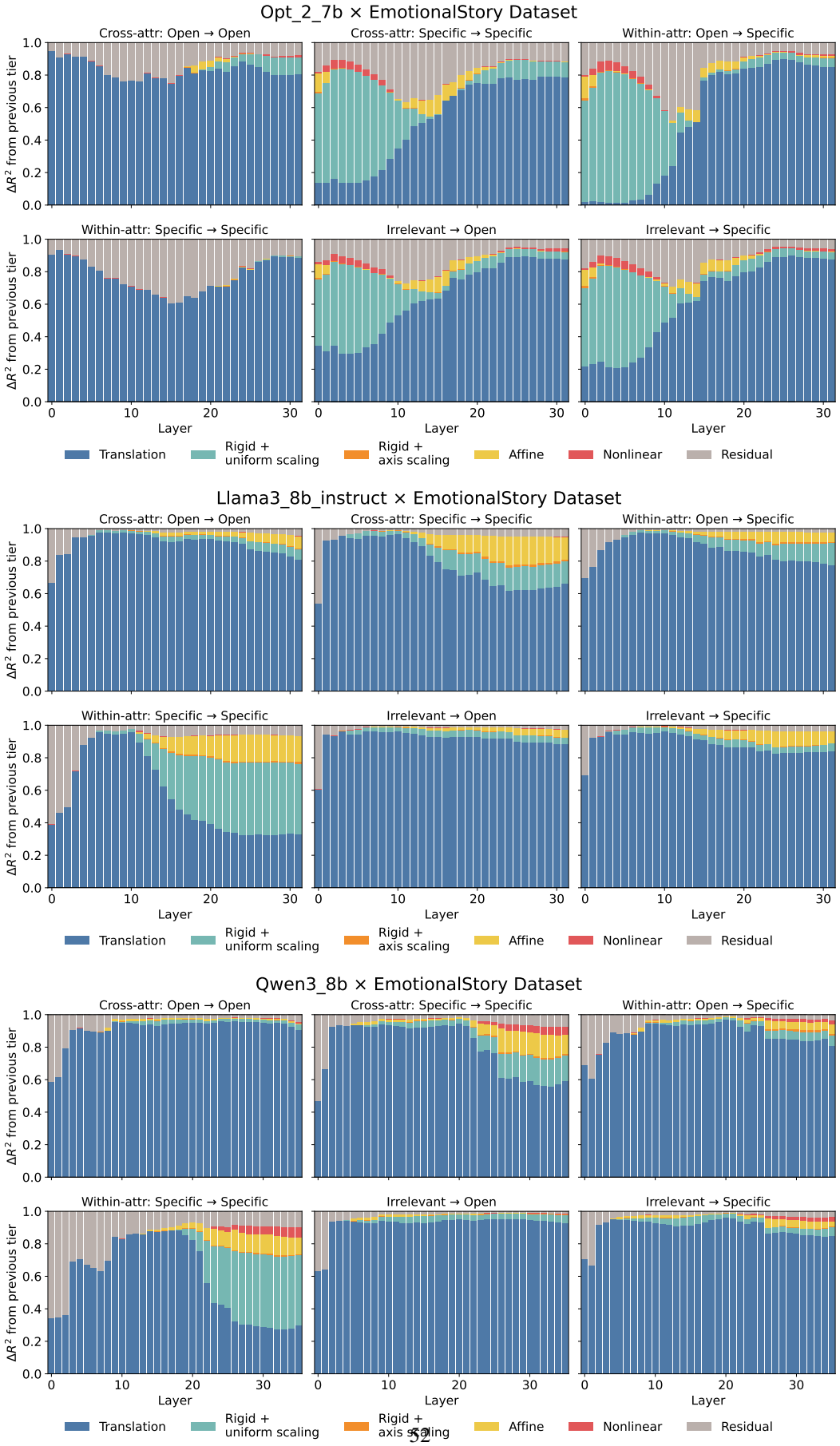

Figure 33: Incremental R^2 of each transformation for EmotionalStory dataset across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

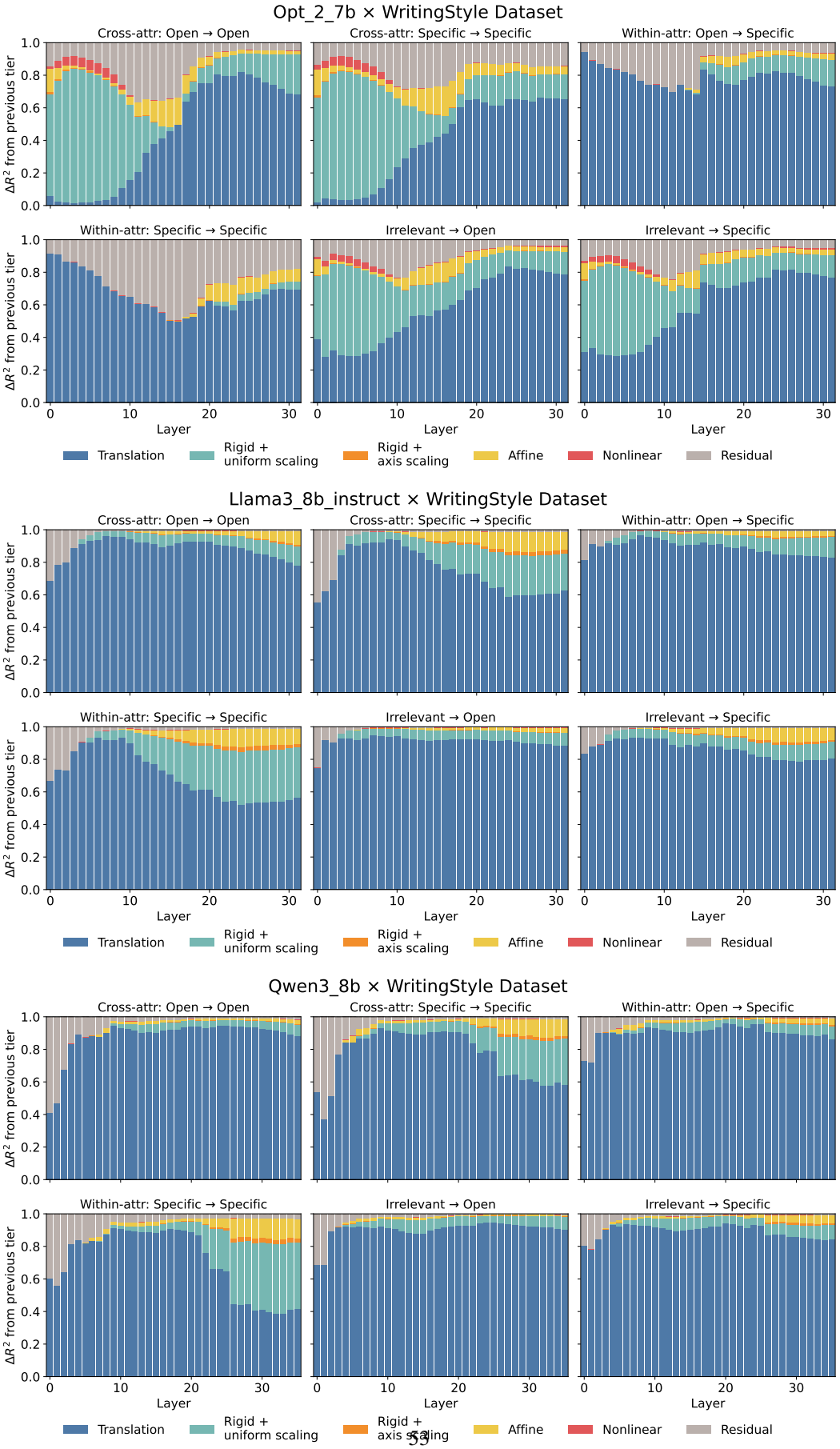

Figure 34: Incremental R^2 of each transformation for WritingStyle across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

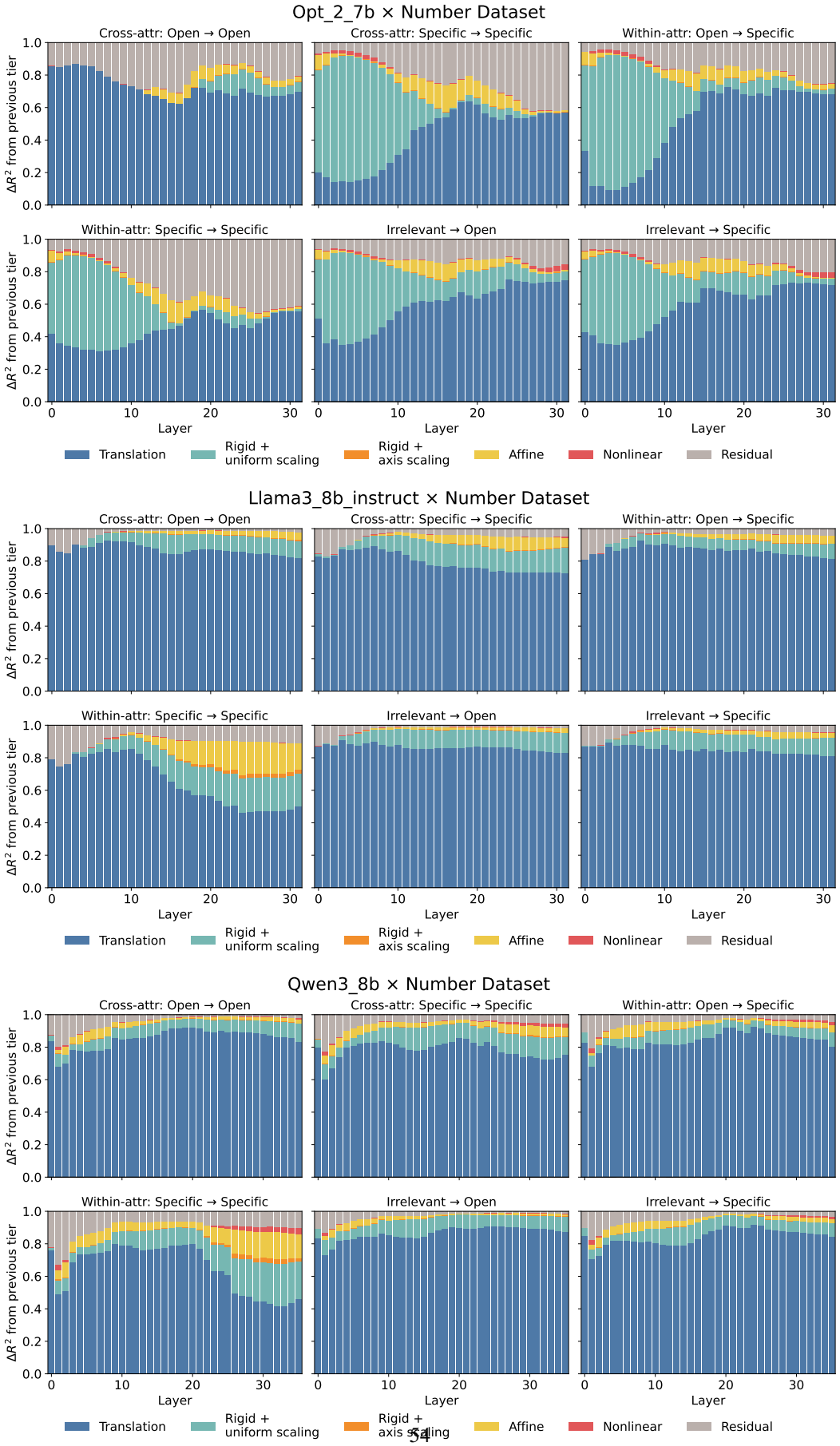

Figure 35: Incremental R^2 of each transformation for Number across layers for OPT-2.7B (top), Llama-3-8B-Instruct (middle), Qwen3-8B (bottom).

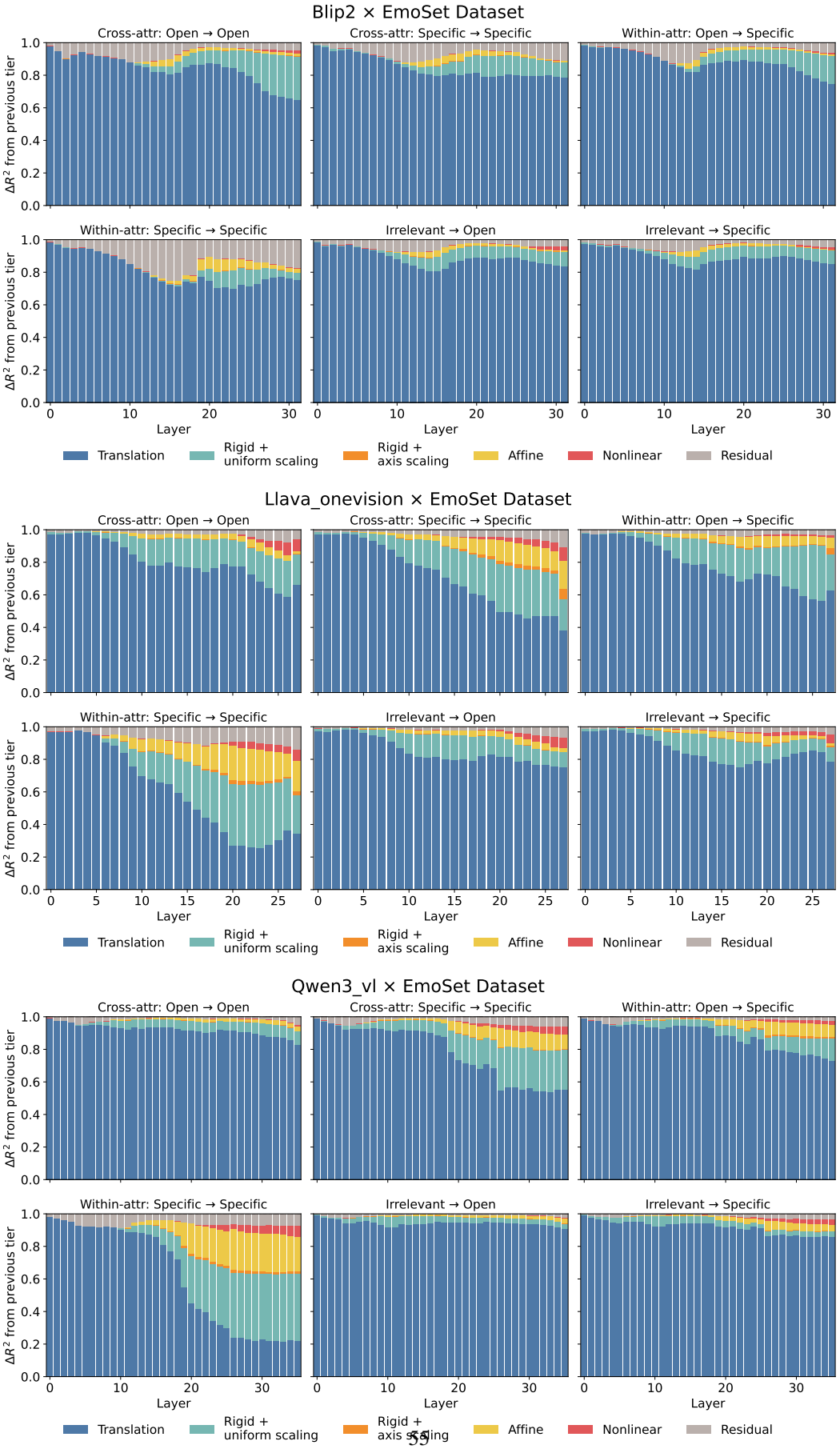

Figure 36: Incremental R^2 of each transformation for EmoSet across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).



Figure 37: Incremental R^2 of each transformation for StyleTransfer across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).

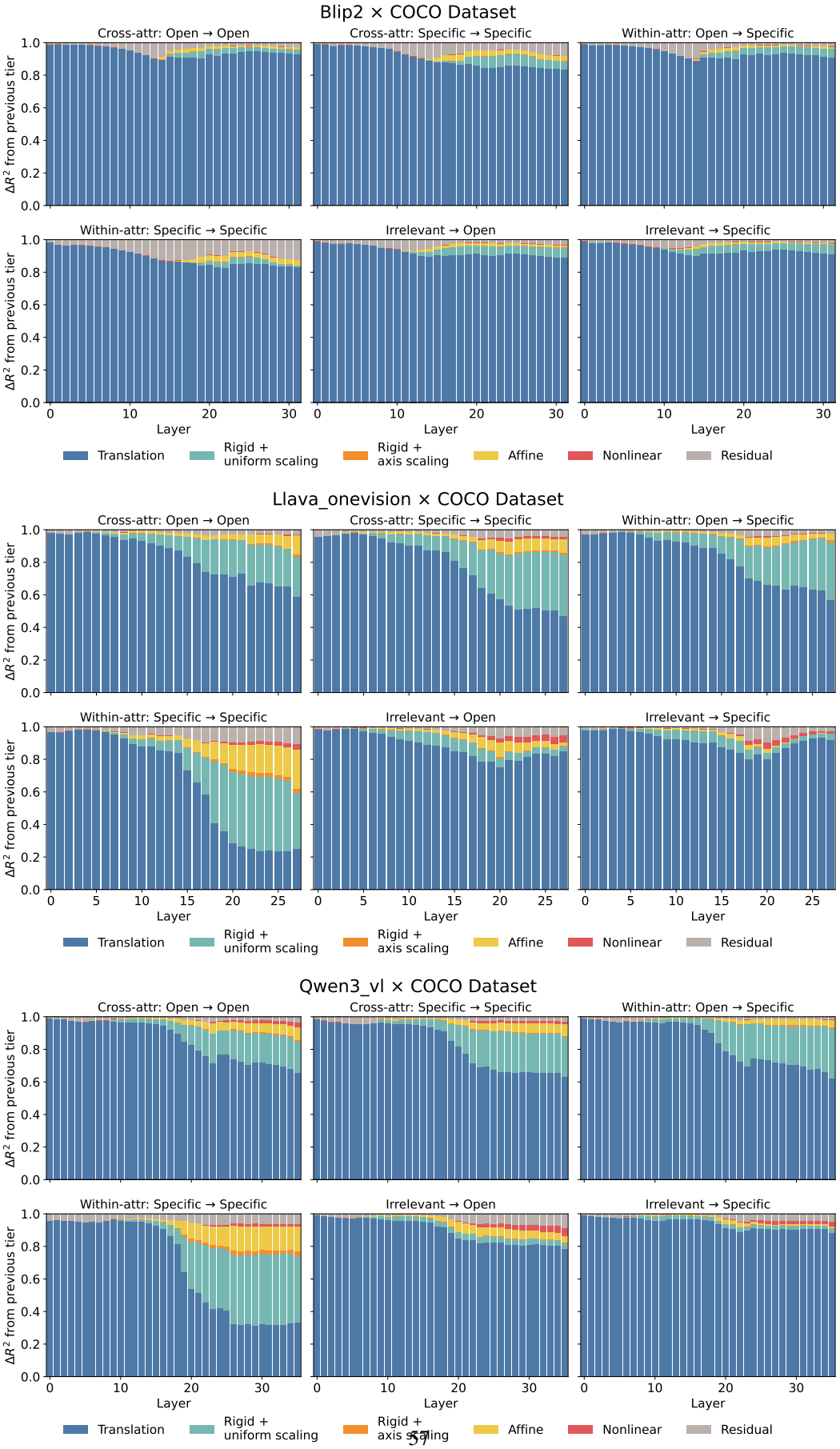

Figure 38: Incremental R^2 of each transformation for COCO across layers for BLIP-2 (top), LLaVA-OneVision-7B (middle), Qwen3-VL-8B (bottom).

M Generalization to prompt paraphrases and out-of-distribution (OOD) datasets

We ask whether the fitted transformations $\widehat{f}_k$ capture task-dependent representational structure that generalizes beyond the specific prompt–stimulus pairings used for fitting. We evaluate two forms of generalization: (i) *prompt paraphrasing*, in which the canonical target prompt B is replaced by semantically equivalent rewordings that query the same attribute, and (ii) *input distribution shift*, in which transformations fitted on one dataset are evaluated on another dataset.

Prompt paraphrasing

Setup. We fix the source prompt A and the canonical target prompt B from the cross-attribute open–open pairing group (G1). We then construct three semantic paraphrases $\{B'_i\}_{i=1}^{3}$ of B, each preserving the queried attribute while varying the surface form of the instruction. For each paraphrase B'_i, we fit each tier-k transformation $\widehat{f}_k^{(i)}$ on the paraphrased prompt pair $(\mathbf{X}^A, \mathbf{X}^{B'_i})$ using the training stimuli, and evaluate the fitted transformation against the canonical target representations $\mathbf{X}^B$ on held-out stimuli. We use the same 5-fold cross-validation protocol as in the main experiments (Section 5).

For paraphrase i, the cumulative paraphrase-generalization score of tier k is

$$R_k^{2,(i)} = \frac{\left\|\mathbf{X}^B - \mathbf{X}^A\right\|_F^2 - \left\|\mathbf{X}^B - \widehat{f}_k^{(i)}(\mathbf{X}^A)\right\|_F^2}{\left\|\mathbf{X}^B - \mathbf{X}^A\right\|_F^2},$$

where both norms are computed over held-out stimuli and scores are averaged across folds. As a reference ceiling, we also fit and evaluate the canonical pair $(\mathbf{X}^A, \mathbf{X}^B)$ under the same cross-validation procedure. If the learned transformation captures the target attribute independently of the exact wording of B, the paraphrase-generalization curves should approach the canonical reference curve.

We evaluate this analysis in one vision-language setting and one language-only setting: LLaVA-OneVision-7B on StyleTransfer, with canonical target prompt $B =$ "What artistic style does the image belong to?" and three paraphrases querying *style*; and Llama3-8B-Instruct on EmotionalStory, with canonical target prompt $B =$ "What emotion does this text express?" and three paraphrases querying *emotion*.

Results. For Llama3-8B-Instruct on EmotionalStory (Fig. 39), the paraphrase-generalization curves closely track the canonical reference curve, except in the earliest layers. This suggests that, across most layers, the fitted transformations capture a prompt-induced representational change that is largely stable across semantically equivalent rewordings of the target prompt. For LLaVA-OneVision-7B on StyleTransfer (Fig. 40), the paraphrase curves approach the canonical reference primarily in deeper layers. This indicates that paraphrase-invariant transformations emerge more strongly at later stages of LLaVA, whereas earlier layers remain more sensitive to the surface form of the prompt.

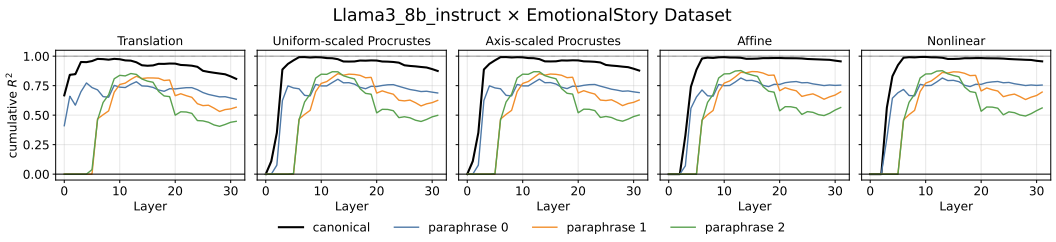

Figure 39: Cumulative cross-validated R^2 of each transformation under prompt paraphrasing, evaluated on the canonical $\mathbf{X}^B$ on held-out stimuli for Llama3-8B-Instruct on EmotionalStory (source A = topic, canonical target B = emotion; three paraphrases of emotion).

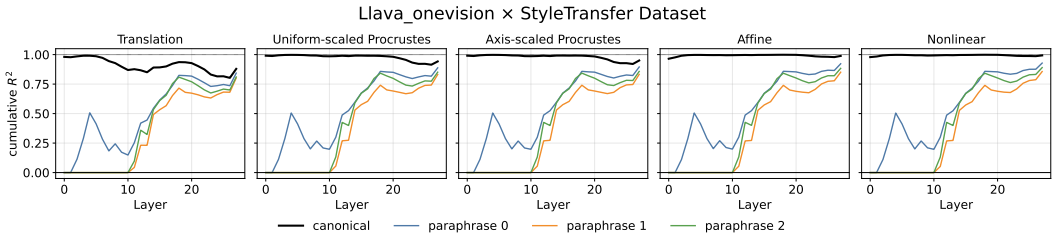

Figure 40: Cumulative cross-validated R^2 of each transformation under prompt paraphrasing, evaluated on the canonical $\mathbf{X}^B$ on held-out stimuli for LLaVA-OneVision-7B on StyleTransfer(source A = scene, canonical target B = style; three paraphrases of style).

Evaluation on out-of-distribution (OOD) datasets

Setup. We next test whether transformations fitted on one stimulus distribution generalize to a distinct dataset from the same modality. For each $(\text{model}, D_{\text{src}}, D_{\text{tgt}})$ triple, we fit each tier-k transformation $\widehat{f}_k^{\text{src}}$ using all stimuli from the source dataset D_{src}. We then apply the fitted transformation to all stimuli in the target dataset D_{tgt} under the same prompt pair and compute the per-tier incremental contribution ΔR_k^2. As in the paraphrase analysis, we evaluate one vision-language setting (LLaVA-OneVision-7B: StyleTransfer $\leftrightarrow$ EmoSet) and one language-only setting (Llama3-8B-Instruct: WritingStyle $\leftrightarrow$ EmotionalStory).

For comparison, the diagonal entries in Fig. 41 show matched in-domain performance estimated by 5-fold cross-validation within each dataset, whereas the off-diagonal entries show cross-dataset transfer from D_{src} to D_{tgt}.

Results. Transformations fitted on one dataset retain substantial explanatory power when transferred to a different dataset from the same modality, but the decomposition across transformation tiers changes under distribution shift (Fig. 41). In the language-only setting, transfer from EmotionalStory to WritingStyle shows the strongest degradation: the translation component decreases markedly in deeper layers relative to the in-domain condition, with the loss primarily redistributed to the affine tier and the residual. The reverse direction, WritingStyle to EmotionalStory, transfers more robustly, suggesting an asymmetry in how dataset-specific structure contributes to the fitted transformation.

In the vision-language setting, the in-domain decompositions are dominated by translation across layers. Under cross-dataset transfer, however, a larger fraction of the explained and unexplained variance is assigned to affine, nonlinear, and residual components. Thus, while a substantial component of the fitted transformation is shared across datasets, the relative contribution of simple translation versus higher-order transformations remains dataset-dependent.

Together with the paraphrase analysis, these results indicate that the nested geometric hierarchy captures prompt-dependent representational structure that is partly invariant to prompt rewording and partly transferable across input distributions. At the same time, the tier-wise allocation of variance is sensitive to the stimulus distribution, suggesting that prompt-induced representational changes contain both task-level and stimulus-distribution-specific components.

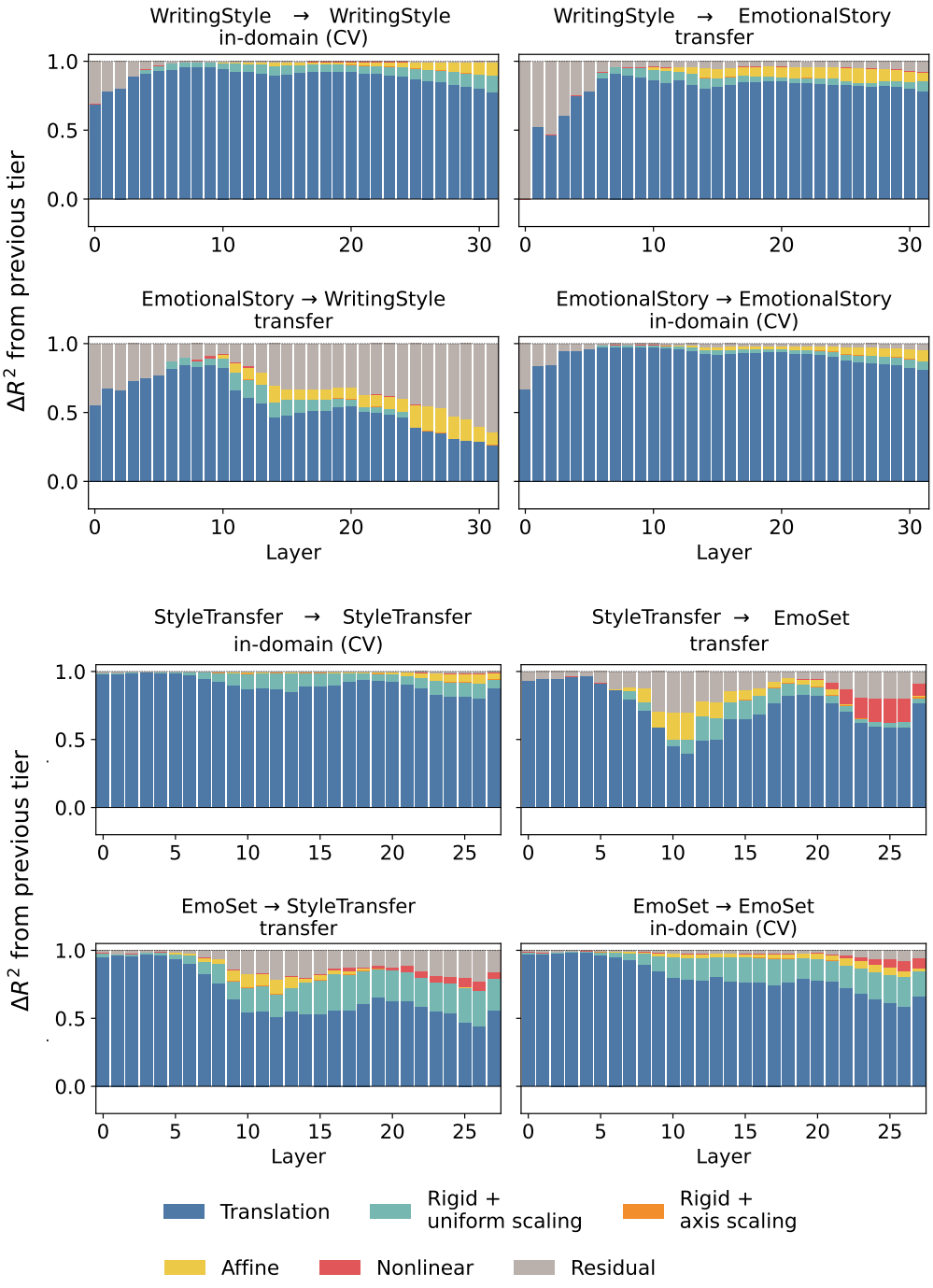

Figure 41: Out-of-distribution (OOD) generalization of the nested geometric decomposition. Top: Llama3-8B-Instruct. Bottom: LLaVA-OneVision-7B. Rows denote the source dataset used to fit the transformation, and columns denote the target dataset used for evaluation. Diagonal panels show matched in-domain 5-fold cross-validation; off-diagonal panels show cross-dataset transfer. Each bar shows the per-layer stacked decomposition of incremental ΔR_k^2 into transformation tiers and residual for cross-attribute open–open prompt pairs.

N Cross-dataset transfer across all six models

Table 10 reports held-out cumulative R^2 for all six models at the five tiers of the nested decomposition: translation, uniform scaling plus rigid, axis scaling plus rigid, affine, and nonlinear. For each model the in-domain row averages the three within-dataset fits, one per source dataset, whereas the cross-dataset row averages the six transfers obtained by fitting on one dataset and evaluating on the other two. Over the six models this gives 18 in-domain fits and 36 cross-dataset transfers. The in-domain row reproduces the per-model values reported in the author response, so the two rows of each block are on a common scale and can be read as a retention comparison.

Table 10: Held-out cumulative R^2 at each tier, averaged within every model. In-domain averages the three within-dataset fits; cross-dataset averages the six transfers to the two held-out datasets. Negative values indicate a fit that performs worse than the direction-symmetric no-transformation baseline.

Model	Condition	Translation	Uniform scaling + rigid	Axis scaling + rigid	Affine	Nonlinear
Llama-3-8B-Instruct	in-domain	0.875	0.945	0.965	0.986	0.984
	cross-dataset	0.561	0.557	0.581	0.713	0.592
Qwen3-8B	in-domain	0.922	0.969	0.977	0.984	0.983
	cross-dataset	0.662	0.716	0.734	0.723	0.710
OPT-2.7B	in-domain	0.749	0.828	0.832	0.874	0.866
	cross-dataset	0.493	−0.283	−0.305	−0.162	−0.272
BLIP-2	in-domain	0.852	0.951	0.955	0.969	0.968
	cross-dataset	0.724	0.841	0.837	0.829	0.824
LLaVA-OneVision	in-domain	0.772	0.929	0.949	0.967	0.967
	cross-dataset	0.379	0.520	0.491	0.524	0.520
Qwen3-VL	in-domain	0.819	0.942	0.962	0.979	0.979
	cross-dataset	0.552	0.842	0.835	0.838	0.837

Nonlinear versus affine retention. Define the fraction retained by a tier as its cross-dataset value divided by the matched in-domain value for the same model on the target dataset. Under this definition the nonlinear tier retains a smaller fraction than the affine tier in 31 of the 36 cross-dataset transfers. In the remaining five transfers the two fractions agree to three decimal places, so the nonlinear tier never retains a clearly larger fraction than the affine tier. Compared in raw held-out cumulative R^2 rather than as a fraction, the nonlinear value lies below the affine value in all 36 transfers. The extra capacity of the nonlinear tier is therefore the part of the fitted map that is most specific to the source stimulus distribution.

Transfers that fail outright. Exactly one of the 36 cross-dataset transfers is negative at every one of the five tiers: LLaVA-OneVision fitted on COCO and evaluated on StyleTransfer. There every tier performs worse than the no-transformation baseline, so the fitted map is actively harmful on the target distribution rather than merely weakened. Five further transfers are negative at some tiers but not all, four of them for OPT-2.7B and one for Llama-3-8B-Instruct; in each of these the translation tier stays positive while one or more of the richer tiers goes negative. This is why the OPT-2.7B cross-dataset row in Table 10 is negative beyond the translation tier despite a positive in-domain profile at every tier: a high in-domain value does not guarantee transfer.

Relation to the per-layer figure. Figure 41 shows the per-layer stacked decomposition for the source-by-target matrix, and the re-centring control that re-estimates only the target centroids while holding the source transformation fixed was run for Llama-3-8B-Instruct and LLaVA-OneVision. That control therefore covers two of the six models, whereas Table 10 aggregates over layers and over all six models with no re-centring applied. The table and that figure are not directly comparable, and the table must not be read as the aggregate of the figure. Reading the two together, the transformation family and several of its geometric components generalize across stimulus distributions, whereas the fitted offset and the higher-capacity refinements are partly specific to the stimulus distribution they were fitted on.

O Pretraining and post-training as a source of the OPT-2.7B profile

Among the six models studied here, OPT-2.7B is the only pretrained-only checkpoint: Llama-3-8B-Instruct and Qwen3-8B have both undergone instruction-oriented post-training. OPT-2.7B is also the language model with the most distinctive tier profile, showing a larger and earlier contribution from rigid transformations than Llama-3-8B-Instruct and Qwen3-8B (Fig. 3). One possible interpretation is that instruction-oriented post-training encourages instructions to be mapped more directly into a shared task-conditioned representational basis, whereas a pretrained-only model may require a greater reorientation of its representations when the instruction changes.

We present this as a hypothesis and not as a causal conclusion. In our model set, post-training status is confounded with architecture, model size, training data, and model family, so the difference in tier profile cannot be attributed to post-training alone. Isolating that effect would require a controlled comparison between matched base and instruction-tuned checkpoints of the same model, which we leave to future work.

P Geometry of the full stimulus span

The decomposition in the main text is fitted at a single token position: the layer-ℓ hidden state at the final input token, which we call the readout position. This appendix asks whether the same geometry also holds at every token of the stimulus span. The multi-token analyses in this section and in Appendices Q and R all use Meta-Llama-3-8B-Instruct on EmotionalStory unless stated otherwise, with the LLM input order instruction, stimulus, readout. The three VLMs are excluded by construction, because their visual stimulus tokens precede the instruction and therefore, under causal attention, their states cannot be influenced by it. At each position of the stimulus span we refit the nested ladder between prompt A and prompt B with the protocol of Section 3, and we additionally compare the prompt-A and prompt-B representational dissimilarity matrices built from the states at that position.

Prompting leaves only a small footprint inside the span. The normalized activation change is 0.29 at the first stimulus token and 0.24 at the last stimulus token, with a median of approximately 0.14 across interior tokens. Across late layers, rigid transformation with uniform scaling ($\mathcal{F}_{\mathrm{O_u}}$) reaches held-out R^2 of 0.97 to 1.00 for most stimulus positions. Table 11 reports the three measured quantities by span position.

Table 11: Geometry across the stimulus span, late layers. Ranges are across layers within each group of positions. The readout token is included as a reference and is not part of the stimulus span.

Span position	Translation R^2	Uniform scaling and rigid R^2	RDM correlation
Interior tokens	0.956–0.969	0.979–0.988	0.972–0.985
First stimulus token	0.854–0.913	0.990–0.996	0.973–0.985
Last stimulus token	0.761–0.772	0.883–0.893	0.877–0.886
Readout token (reference)	0.189–0.241	0.532–0.573	0.469–0.516

Two readings follow. The stimulus span is well described by the same low-complexity maps that the main text fits, and translation alone already accounts for most of the prompt-induced change at interior tokens, whereas the readout position is by far the hardest position to fit on all three measures. Multidimensional scaling across datasets agrees: under the two prompts the geometry of the stimulus tokens either aligns with one prompt's task structure for both prompts or stays unaligned with either, so prompting largely preserves stimulus-borne geometry rather than creating a new task-specific geometry inside the stimulus tokens.

Q Multi-token and multi-layer interventions

The model, dataset, and input layout are those of Appendix P. Here we ask where the prompt effect is written and where it is read, by replacing prompt-A states with the corresponding prompt-B states at whole groups of positions rather than at the readout position alone. Accuracy is scored on prompt B's target attribute against the dataset ground truth. In Table 12 the single-layer conditions sweep

every layer from 1 to 32 and report the range over that sweep, whereas the range conditions patch all layers in the stated band simultaneously.

Replacing the stimulus span has essentially no behavioral effect: the single-layer sweep spans 0.132 to 0.158 against an unpatched prompt-A baseline of 0.132. Adding the readout position lifts the best single layer to 0.213, and patching that pair of positions across all 32 layers reaches 0.267, still far below the end-to-end prompt-B reference of 0.661. The positive control that patches the instruction span instead recovers 0.657 over layers 1 to 7 and 0.660 over all 32 layers, both at the end-to-end level. Dropping the final instruction token from that control does not reduce the effect: 0.660 and 0.665 for the same two bands.

Table 12: Positional causal patching, accuracy on the prompt-B attribute. Single-layer conditions give the range over a sweep of all 32 layers; paired values give the two patched layer bands, L1–7 and L1–32.

Condition	Accuracy
Prompt A, no patch	0.132
Stimulus span patched at one layer (L1–32)	0.132–0.158, peak at layer 12
Stimulus span and readout patched at one layer (L1–32)	0.134–0.213, peak at layer 12
Stimulus span and readout patched across layers (L1–7 / L1–32)	0.147 / 0.267
Positive control: instruction span across layers (L1–7 / L1–32)	0.657 / 0.660
Positive control without final instruction token (L1–7 / L1–32)	0.660 / 0.665
Prompt B, end to end	0.661

Writing happens at the instruction, and reading at the readout position. The instruction positions carry the prompt effect in full, and they already carry it within the first seven layers, since extending the band to all 32 layers gives no further gain. That effect is distributed and redundant across the span rather than localized at its end, because excluding the final instruction token leaves it intact. The stimulus tokens are not where the effect lives: patching them moves behavior only once the readout position is included, which is what one expects if the readout position first reads the instruction and then uses it to select task-relevant information from the stimulus.

R What the readout position attends to

The setup is again the one of Appendix P. Rather than intervening, we read the attention of the readout position directly and ask which input positions it draws on at each depth. For this input the uniform attention share is $1/138 \approx 0.007$, and it is the reference against which every value in Table 13 should be read.

Table 13: Attention from the readout position, by layer band. Uniform share $1/138 \approx 0.007$. Ranges are across the layers in the band.

Layers	Readout to final instruction token	Readout to stimulus span, summed
L1–3	0.003–0.006	0.546, 0.101, 0.037 at layers 1, 2, 3
L4–6	rising, 0.010–0.063	0.060–0.096
L7–8	**peak, 0.093–0.107 (13–15$\times$ uniform)**	0.052–0.071
L9–12	0.044–0.099	0.049–0.131
L13–15	0.036–0.075	**peak, 0.126–0.189**
L16–23	0.010–0.037	0.027–0.093
L24–32	0.002–0.007	0.019–0.052

Attention to the final instruction token rises through the early layers and peaks over layers 7 and 8 at 0.093 to 0.107, which is 13 to 15 times the uniform share, and that token takes the largest single share of any instruction position there. At layer 7 the readout position assigns 0.212 total attention to the instruction span. Attention to the stimulus span peaks later, over layers 13 to 15 at 0.126 to 0.189, following an isolated layer-1 value of 0.546 that drops to 0.101 and 0.037 at layers 2 and 3. Patching only the final instruction token localizes the read in depth: against a baseline of 0.154 in this

sub-analysis, layer 1 has no effect, layer 8 raises accuracy to 0.214, layer 14 has no effect despite attention still well above uniform, and layer 32 sits exactly at baseline.

The depth ordering is a read of the instruction before a read of the stimulus, which is the pattern the interventions in Appendix Q predict. The final instruction token is the strongest single read site, and layer 8 is the depth at which reading it matters behaviorally. It is nevertheless not a bottleneck: attention well above uniform at layer 14 carries no causal effect, and removing the token from the instruction-span control of Appendix Q leaves the behavioral effect intact.

S Dimensionality of Procustes alignment

Fig. 4**b** (bottom) shows the alignment dimensionality using a 1% relative threshold on the singular values of cross-covariance $\widetilde{\mathbf{X}}^{A\top}\widetilde{\mathbf{X}}^{B}$ ($\sigma_i > 0.01\,\sigma_{\max}$). For completeness we also report the stricter 10% threshold ($\sigma_i > 0.10\,\sigma_{\max}$) in Fig. 42 for LLMs.

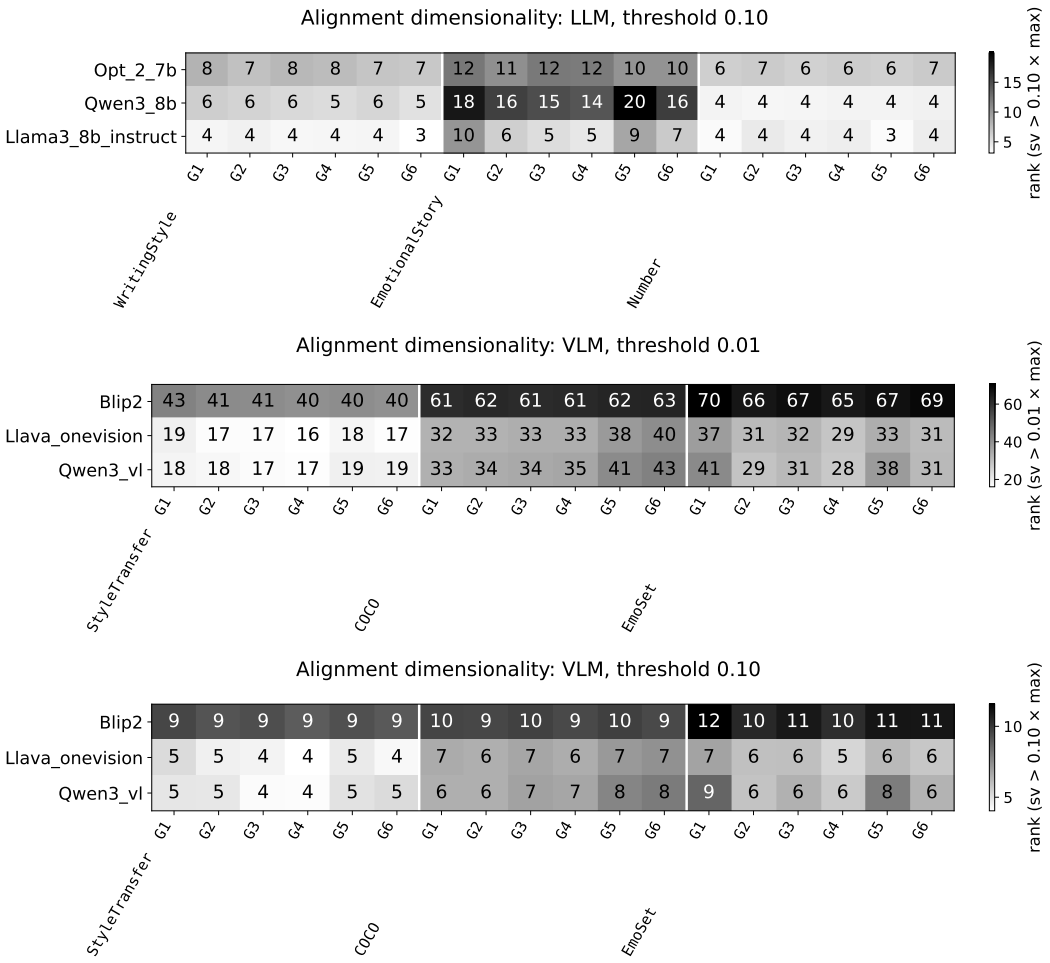

Figure 42: Alignment dimensionality. (top) the centered cross-prompt cross-covariance at the stricter 10%-of-$\sigma_{\max}$ threshold (LLM, all six prompt-pair groups). (bottom and middke) llm. two thresholds. the layout across (model $\times$ dataset $\times$ group) is the same as in Fig. 4**b**.

T Example generated texts after interventions

We provide qualitative examples of model generations after replacing an internal representation with the output of each fitted transformation tier. For a stimulus s_i and layer ℓ, we run the model with prompt A, replace the residual-stream activation at the final input-prompt token by $\widehat{f}_k(\Phi^A(s_i))$, and then greedily decode 50–100 tokens from the modified state. We compare seven conditions: **prompt** A without intervention; **translation**; **rigid_uni**, corresponding to an orthogonal transformation plus uniform scaling; **rigid_axis**, corresponding to an orthogonal transformation plus axis-wise scaling; **affine**; **nonlinear**; and **prompt** B, an oracle activation reference in which the prompt-A residual stream at the same layer and token is replaced by the held-out representation $\Phi^B(s_i)$. This oracle condition preserves the remaining prompt-A context and is therefore distinct from running the model end-to-end under prompt B.

G1 / LLaVA-OneVision-7B $\times$ COCO. Prompt A: *"Are there people in this image?"* Prompt$B$: *"How many people are in this image?"* This prompt pair tests whether an intervention can shift the model from binary detection to numerical counting. At the mid-depth layer $\ell = 22$, higher-tier transformations replace the prompt-A yes/no response format with count-like answers (Table 14).

Table 14: G1 / LLaVA-OneVision-7B $\times$ COCO, prompt pair *detect* $\rightarrow$ *count* at $\ell = 22$.

Stim	prompt A	translation	rigid_uni	rigid_axis	affine	nonlinear	prompt B
8	yes	two	two	two	two	two	two
12	no	no	0	0	0	0	0
13	no	no	0	0	zero	0	zero
15	yes	five	five	five	five	five	five
25	yes	four	four	four	four	four	four
29	yes	two	two	two	10	10	10
41	yes	one	one	one	one	one	one

G1 / Qwen3-VL-8B $\times$ COCO. We use the same detection-to-counting prompt pair and apply interventions

at layer $\ell = 27$ (Table 15). Because

Qwen3-VL often produces longer descriptive responses, we report only the

first sentence in each cell. Translation and the orthogonal-plus-uniform

scaling tier largely preserve the detection framing of prompt A

(e.g., "Yes, there are people . . ."), whereas affine and nonlinear

interventions more consistently shift the output toward a numerical

count.

Table 15: G1 / Qwen3-VL-8B $\times$ COCO, prompt pair *detect* $\rightarrow$ *count* at $\ell = 27$ (first sentence shown per cell). Cells with identical text within a row are grouped.

Stim	Generated text per level group (groups separated by "|")
25	*prompt* A: "Yes, there are people in this image." | *T, O_u, O_a*: "Based on the image provided, yes, there are people in this image." | *L, N, prompt* B: "There are four people in this image."
69	*prompt* A: "Yes, there are people in this image." | *T, O_u*: "Based on the image provided, yes, there are people in this image." | *O_a, L, N, prompt* B: "There are two people in the image."
94	*prompt* A: "Yes, there are two people in this image." | *T*: "Based on the visual information in the image, yes, there are people in this image." | *O_u, O_a, L, N, prompt* B: "There are two people in this image."
211	*prompt* A: "Yes, there are people in this image." | *T, O_u, O_a*: "Based on the image provided, yes, there are people in this image." | *L, N, prompt* B: "There are three people in this image."
237	*prompt* A: "Yes, there are people in this image." | *T*: "Based on the image provided, yes, there are people in this image." | *O_u, O_a, L, N, prompt* B: "There are five people in this image."

G5 / LLaVA-OneVision-7B $\times$ COCO: irrelevant "capital" prompt. Prompt A: *"What is the capital of France?"* Prompt B: *"How many people are in this image?"* This prompt pair tests whether an intervention can induce an image-dependent counting response even when the source prompt is task-irrelevant. Without intervention, the model answers the source question with "Paris." After applying intermediate- and higher-tier transformations, the generated response shifts toward the count requested by prompt B (Table 16). This suggests that the fitted transformations can inject target-task structure beyond the semantic content explicitly requested by the source prompt.

Table 16: G5 / LLaVA-OneVision-7B $\times$ COCO, prompt pair *capital* $\rightarrow$ *count* at $\ell = 22$.

Stim	prompt A	translation	rigid_uni	rigid_axis	affine	nonlinear	prompt B
8	paris	two	two	two	two	two	two
12	paris	two	zero	zero	0	0	0
13	paris	two	two	two	0	zero	zero
15	paris	two	five	five	five	five	five
25	paris	two	four	four	four	four	four
29	paris	two	two	two	two	two	10
41	paris	two	One	One	three	one	one

G5 / LLaVA-OneVision-7B $\times$ COCO: irrelevant "arithmetic" prompt. Prompt A: *"What is 2+2?"* Prompt B: *"How many people are in this image?"* This condition provides a second task-irrelevant source prompt. The unmodified model produces the arithmetic answer "4," whereas interventions can shift the response toward an image-dependent count (Table 17).

Table 17: G5 / LLaVA-OneVision-7B $\times$ COCO, prompt pair *arithmetic* $\rightarrow$ *count* at $\ell = 22$.

Stim	prompt A	translation	rigid_uni	rigid_axis	affine	nonlinear	prompt B
8	4	two	two	two	two	two	two
12	4	two	zero	zero	0	0	0
13	4	two	0	0	0	0	0
15	4	two	five	five	five	five	five
25	4	4	4	4	4	4	4
29	4	two	two	two	two	three	10
41	4	two	1	two	1	2	1

U Asset licenses and credits

Pretrained models. All models used in this paper are publicly available HuggingFace checkpoints, cited via their original papers and listed with the URL of the specific revision and the license under which we used them:

- LLaVA-OneVision-7B [Li et al., 2024] – `https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf` – Apache 2.0.
- Qwen3-VL-8B-Instruct [Bai et al., 2025] – `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct` – Apache 2.0.
- Qwen3-8B [Yang et al., 2025] – `https://huggingface.co/Qwen/Qwen3-8B` – Apache 2.0.
- BLIP-2 (OPT-2.7B backbone) [Li et al., 2023a] – `https://huggingface.co/Salesforce/blip2-opt-2.7b` – BSD-3-Clause.
- Meta-Llama-3-8B-Instruct [Grattafiori et al., 2024] – `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct` – Meta Llama 3 Community License Agreement.
- OPT-2.7B [Zhang et al., 2022] – `https://huggingface.co/facebook/opt-2.7b` – OPT-175B License Agreement (research use; the same agreement covers all OPT model sizes).

Image datasets.

- COCO val2017 [Lin et al., 2014] – annotations released under CC BY 4.0; the underlying images are subject to Flickr Terms of Use. We use a 1,000-image subset balanced by supercategory.
- EmoSet [Yang et al., 2023] – non-commercial research use only, no redistribution permitted; used in compliance with the dataset's terms of use. We use a 1,600-image 8×4 emotion-by-content subset.
- StyleTransfer [Boger and Firestone, 2025] – 1,920 photographs rendered in seven styles, made available alongside the published paper in *Nature Human Behaviour* 9 (2025) 2497–2509.

Curated text stimulus sets. The three LLM prompt-pair sets (EMOTIONALSTORY, WRITINGSTYLE, NUMBER) were authored by us. They consist of factual, style, and numeric prompts and contain no personal or sensitive content. They will be released alongside the analysis code under CC BY 4.0.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state the contributions (the nested geometric decomposition $\mathcal{F}_{\mathrm{T}} \subset \ldots \subset \mathcal{F}_{\mathrm{N}}$, the empirical finding that low-complexity linear maps recover most of natural prompting's effect, and the causal-intervention validation across six models and six datasets); these match the experimental results in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A Limitations and broader-impacts paragraph at the end of Sec. 7 flags (i) the zero-shot instruction scope, few-shot demonstrations, soft prompts, and longer context-conditioning regimes are untested; (ii) the single-layer, last-token granularity of the causal intervention; and (iii) that out-of-distribution evaluation of the fitted maps needs more systematic experiments in the future.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical. The nested R^2 decomposition uses standard linear-algebraic tools (mean shift, orthogonal Procrustes for the rigid+uniform-scale tier, sequential per-axis least squares for the rigid+axis-scaling tier, ridge regression for the affine tier, and a small shared MLP fit by mini-batch AdamW for the nonlinear tier); no novel theorems are proved. Estimator definitions, the cumulative/incremental R^2 partition-of-unity identity, and the nested-chain inclusions are given in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 3–5 and Appendices A, E, F fully specify the model checkpoints, dataset subsets, cross-validation protocol, intervention level definitions, and pipeline parameters needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All image datasets (COCO, EmoSet, StyleTransfer) are publicly available; upon de-anonymization, with documentation in Appendices E and U and a README in the released repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation details are specified in Sec. 3 and Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report individual results in Appendix while showing pooled results with error bar or band in main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All pair pipelines ran on a single internal-cluster GPU per pair (A40/A100-class) coordinated via Slurm arrays. Total compute is approximately 15,000 GPU-hours: $\approx 14{,}000$ for the main task pipeline ($\sim$11.5 h $\times$ 1,213 prompt pairs across 15 model$\times$dataset cells), ≈ 500 for the exploratory phase. Evaluation jobs are CPU-only and negligible by comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly released models and datasets, performs read-only analysis of internal representations, and involves no human subjects, deployment, or sensitive data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussed in the closing paragraph of Sec. 7: the framework contributes to interpretability/transparency of prompting (positive); the same understanding could marginally inform more efficient activation-level steering, including along unsafe directions, but the marginal contribution to misuse risk is small relative to existing activation-steering literature.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No high-risk-of-misuse asset is released. The pretrained models are public HuggingFace checkpoints used under their original licenses (Appendix U); the only new artefacts are author-curated factual/style/numeric prompt sets and analysis code, which do not pose misuse risk beyond that of the underlying public models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All pretrained models (LLaVA-OneVision, Qwen3-VL, Qwen3, BLIP-2, Llama-3, OPT) and image datasets (COCO, EmoSet, StyleTransfer) are credited and listed with their respective licenses in Appendix U.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets`

has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are the curated LLM prompt-pair sets (EMOTIONALSTORY, WRITINGSTYLE, NUMBER) and the analysis code. The prompt sets are documented in Appendix E; the code repository will include a README with run instructions and the licensing information summarized in Appendix U.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subject research is involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: A separate, frontier LLM (Anthropic Claude Sonnet) was used to render the synthetic text datasets EmotionalStory and WritingStyle; the stimulus-generation procedure is documented in Appendix E and the generation scripts will be released with the rest of the analysis code.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.