Real Images, Worse Judgments: Evaluating Vision-Language Models on Concreteness and Imagery

Anonymous ACL submission

Abstract

Visual grounding is often assumed to benefit from incorporating visual information into language understanding. In this work, we critically examine this assumption by evaluating the ability of state-of-the-art vision–language models to capture human-like judgments of word concreteness and imagery. These judgments reflect the extent to which words evoke sensory experiences and have clear visual referents, making them a natural testbed for assessing visual grounding capabilities. Surprisingly, we find that adding real images does not yield consistent gains and often hurts the alignment with human ratings, especially for abstract and low-imagery words. Through comprehensive analyses, including probing, canonical correlation analysis, and interpretability techniques, we find that real images are associated with representational shifts and greater reliance on potentially spurious visual cues that can interfere with linguistic encoding. To address this issue, we demonstrate that a simple inference-time intervention, namely, augmenting prompts with explicit instructions to focus solely on textual content, can mitigate real-image-associated performance degradation, particularly for abstract and low-imagery words. Our findings highlight a critical vulnerability in current vision–language models and underscore the need for more robust mechanisms to guide visual grounding in alignment with human cognition.

1 Introduction

Visual grounding, the task of linking linguistic elements to visual representations, has gained renewed attention with the rise of multimodal models (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023). By incorporating visual input, these models are often expected to enrich linguistic understanding with concrete sensory context, and they have shown strong performance on tasks such as image captioning (Li et al., 2023a), visual question answering (Alayrac et al., 2022), and multimodal reasoning (Liu et al., 2023).

However, current multimodal models still face challenges in grounding language in multimodal contexts consistent with human cognition. Prior work shows that they can over-rely on superficial correlations in the training data (Goyal et al., 2017; Agrawal et al., 2018; Yüksekgönül et al., 2023), leading for example to object hallucinations or errors in spatial interpretation driven by linguistic co-occurrence biases (Li et al., 2023b). These failure modes suggest that strong downstream performance does not necessarily imply robust grounding, and they raise broader concerns about whether apparent visual grounding reflects genuine cross-modal understanding or brittle shortcuts.

In this work, we investigate the extent to which state-of-the-art vision–language models (VLMs) can capture human judgments of concreteness and imagery, two psycholinguistic properties closely tied to perceptual grounding. Concrete and highly imaginable words tend to have clearer visual referents than abstract and low-imagery words, making these ratings a natural testbed for lexical grounding. By examining how well VLMs can capture these aspects of language, we aim to identify specific failure modes and biases that limit their visual grounding capabilities. In summary, our contributions are as follows:

- We evaluate instruction-tuned text-only and vision–language models on human judgments of concreteness and imagery.
- We show that real images can hurt alignment with human ratings, especially for abstract and low-imagery words, and analyze this effect with probing, canonical correlation analysis, and interpretability techniques.
- We demonstrate a simple inference-time prompting intervention that mitigates some of the degradation caused by real-image inputs.

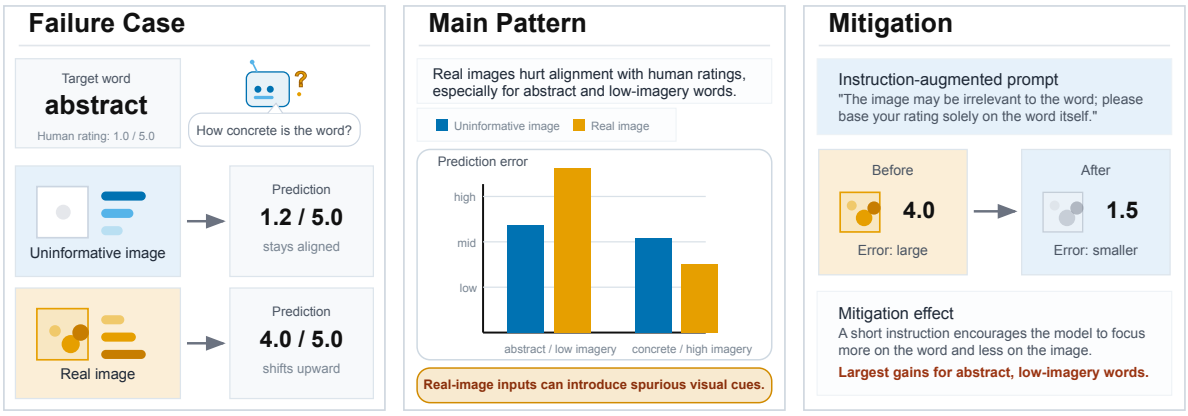


Figure 1: Real images can hurt alignment with human concreteness and imagery judgments, especially for abstract and low-imagery words. This effect is associated with greater reliance on spurious visual cues, and instruction-augmented prompting mitigates it by encouraging the model to base its rating on the word itself.

2 Related Work

Concreteness and imagery. Concreteness and imagery are well-established psycholinguistic properties that quantify the degree to which words are tied to perceptual experience. Concreteness measures how directly a word denotes something perceptible, whereas imagery captures the ease with which a word evokes a mental image (Paivio et al., 1968; Friendly et al., 1982; Turney et al., 2011; Brysbaert et al., 2014). In empirical research, they are typically operationalized through large-scale human rating norms.

Several resources have been developed to measure human judgments of concreteness and imagery. Standard resources include MT40k (Brysbaert et al., 2014), PYM (Paivio et al., 1968) and its extension (Clark and Paivio, 2004), and the Toronto Word Pool (Friendly et al., 1982), which have become widely used benchmarks for studying the relationship between language and perception.

Concreteness and imagery have also been employed as diagnostic signals in AI grounding research. More recently, concreteness prediction has been proposed as a test of visual language understanding, reflecting the extent to which models encode perceptually grounded aspects of meaning (Shi et al., 2019; Alper et al., 2023; Ma et al., 2023, *inter alia*). Human concreteness and imagery ratings therefore offer a principled diagnostic of how well multimodal representations encode perceptually grounded meaning. Motivated by prior work, in this paper, we use these concreteness and imagery ratings as probes of grounding quality in modern multimodal models.

Visual grounding in multimodal models. VLMs learn joint visual–textual representations through large-scale pretraining, ranging from early contrastive models such as CLIP (Radford et al., 2021) to more recent instruction-tuned systems like LLaVA (Liu et al., 2023). A broad line of work shows that incorporating perceptual input improves semantic representations compared to text-only models, particularly for concrete concepts (Hill and Korhonen, 2014; Zhuang et al., 2024b,a). In grounded models, semantic spaces often exhibit a principal dimension aligned with the concrete–abstract continuum, suggesting that visual grounding systematically organizes lexical meaning along perceptual axes. Importantly, prior studies consistently report graded effects: visual information benefits concrete words more strongly than abstract ones, although gains are generally positive across the spectrum.

Building on this line of work, we examine the relationship between visual grounding and human judgments of concreteness and imagery, with a focus on modern instruction-tuned VLMs. In contrast to prior findings, we show that VLMs can underperform text-only models on inference-time judgments of concreteness and imagery in certain conditions.

3 Experiment Setup

3.1 Models

For text-only models, we evaluate LLaMA 2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Qwen2.5 (Yang et al., 2024). For VLMs, we evaluate LLaVA 1.5 (Liu et al., 2024), InternVL3 (Zhu et al., 2025), Pixtral (Agrawal et al., 2024),

	None	White	Noise	ImageNet	Wikimedia
RMSE (All words)					
LLaMA 2	1.15	–	–	–	–
Mistral	0.803	–	–	–	–
Qwen2.5	0.745	–	–	–	–
Gemma 3	**0.745**	0.933	1.185	0.786	0.788
InternVL3	0.883	0.996	1.700	**0.799**	0.813
LLaVA 1.5	1.377	**1.356**	1.358	1.357	1.357
Pixtral	1.138	**0.925**	1.105	0.997	0.961
Qwen2.5-VL	0.930	1.063	0.989	0.945	**0.907**
RMSE (Concreteness < 3)					
LLaMA 2	1.499	–	–	–	–
Mistral	0.939	–	–	–	–
Qwen2.5	0.715	–	–	–	–
Gemma 3	0.730	**0.510**	0.610	0.951	0.926
InternVL3	0.600	0.515	**0.507**	0.837	0.813
LLaVA 1.5	1.908	**1.865**	1.867	1.867	1.867
Pixtral	1.560	**0.696**	1.361	1.303	1.253
Qwen2.5-VL	**0.787**	0.791	0.802	1.092	1.039
RMSE (Concreteness >= 3)					
LLaMA 2	0.738	–	–	–	–
Mistral	0.667	–	–	–	–
Qwen2.5	0.769	–	–	–	–
Gemma 3	0.758	1.178	1.509	**0.614**	0.650
InternVL3	1.066	1.267	2.262	**0.766**	0.813
LLaVA 1.5	**0.649**	0.672	0.674	0.671	0.671
Pixtral	**0.578**	1.082	0.830	0.630	0.612
Qwen2.5-VL	1.035	1.247	1.123	0.799	**0.777**

MT40k (Concreteness)

	None	White	Noise	ImageNet	Wikimedia
RMSE (All words)					
LLaMA 2	1.557	–	–	–	–
Mistral	1.392	–	–	–	–
Qwen2.5	1.275	–	–	–	–
Gemma 3	**1.396**	1.420	1.920	2.051	1.971
InternVL3	1.289	**1.234**	2.100	1.464	1.417
LLaVA 1.5	**1.643**	1.941	2.247	2.267	2.264
Pixtral	2.009	**1.403**	1.577	2.039	2.069
Qwen2.5-VL	**1.385**	1.440	1.590	1.685	1.704
RMSE (Imagery < 4)					
LLaMA 2	1.356	–	–	–	–
Mistral	1.819	–	–	–	–
Qwen2.5	1.389	–	–	–	–
Gemma 3	1.662	1.611	**0.808**	2.680	2.564
InternVL3	1.590	1.372	**0.739**	1.926	1.842
LLaVA 1.5	**2.031**	2.627	3.040	3.065	3.063
Pixtral	2.680	**1.248**	1.867	2.666	2.709
Qwen2.5-VL	**1.628**	1.649	1.799	2.219	2.238
RMSE (Imagery >= 4)					
LLaMA 2	1.725	–	–	–	–
Mistral	0.809	–	–	–	–
Qwen2.5	1.158	–	–	–	–
Gemma 3	**1.087**	1.213	2.558	1.182	1.163
InternVL3	0.918	1.088	2.838	**0.815**	0.836
LLaVA 1.5	1.163	**0.903**	1.045	1.062	1.057
Pixtral	**1.034**	1.534	1.243	1.173	1.182
Qwen2.5-VL	1.109	1.211	1.362	**0.934**	0.959

CP2004B (Imagery)

Table 1: Root mean square error (RMSE) of model predictions compared to human ratings for concreteness (left) and imagery (right). For VLMs, *None* indicates no image input, *White* indicates a white image, *Noise* indicates a random-noise image, *ImageNet* indicates retrieved images from ImageNet, and *Wikimedia* indicates retrieved images from Wikimedia. **Boldface** and underlining indicate the best (lowest RMSE) and worst (highest RMSE) performance, respectively, for each model across visual conditions.

Qwen2.5-VL (Bai et al., 2025), and Gemma 3 (Kamath et al., 2025). All models are instruction-tuned. Within each model family, the text-only model serves as the backbone for its VLM counterpart, e.g., Qwen2.5-VL is the vision–language version of Qwen2.5, which allows for direct comparisons to isolate the effect of vision fine-tuning. Checkpoint and hardware details are provided in Appendices A.1 and A.9.

3.2 Dataset

We evaluate the models on two standard datasets of human ratings: MT40k (Brysbaert et al., 2014) for concreteness and CP2004B (Clark and Paivio, 2004) for imagery. To avoid ambiguities associated with multi-word expressions, we restrict the analysis to single-word entries and evaluate on the shared vocabulary intersection between the two resources, yielding 2,082 words. In MT40k, concreteness is rated on a 5-point scale (1 = highly abstract; 5 = highly concrete). In CP2004B, imagery is rated on a 7-point scale (1 = low imagery; 7 = high imagery). For each word, we retrieve real images from ImageNet (Deng et al., 2009) and Wikimedia Commons using Navigu (Barthel et al., 2023). Because abstract and low-imagery words often lack clear visual referents, these retrieved images should be understood as best-effort matches rather than guaranteed semantic ground truth. In addition to these two real-image sources, we evaluate three controls: a no-image condition (*None*), a white image (*White*), and a random-noise image (*Noise*). Additional dataset preprocessing and visual-input details are provided in Appendices A.2 and A.4.

3.3 Prompt Design

For both tasks, we prompt the models to assign a numerical rating to a target word using the original dataset scales. To reduce prompt sensitivity, we use five prompt variants per task and average the parsed predictions across them. VLMs receive the same textual prompts together with one of the five visual conditions (*None*, *White*, *Noise*, ImageNet, Wikimedia). The full prompt templates are listed in Appendix A.3.

3.4 Evaluation Methodology

Model predictions. For each model and each word, we extract a numerical prediction from the model output and evaluate it against the human rating using root mean square error (RMSE). We report RMSE on the full evaluation set and on abstract/concrete and low/high-imagery subsets. Lower RMSE indicates better alignment with human judgments. Parsing and aggregation details are provided in Appendix A.5.

Probing. We probe layer-wise hidden representations by training regressors to predict concreteness and imagery ratings from them. This lets us test whether different visual conditions strengthen or weaken the encoding of these properties across the network depth. As a sanity check, we also report permuted-label probing baselines in Appendix B, obtained by training the same probes after permuting the target ratings. Complete probing configurations are provided in Appendix A.6.

Canonical correlation analysis. We use canonical correlation analysis (CCA) to compare final-layer hidden representations produced under different visual conditions, such as *None* versus real-image conditions. Given paired representation matrices $\mathbf{X}$ and $\mathbf{Z}$, where rows correspond to words and columns to hidden dimensions under two visual conditions, CCA finds projection vectors $\mathbf{a}$ and $\mathbf{b}$ that maximize

$$\rho = \max_{\mathbf{a},\mathbf{b}} \operatorname{corr}(\mathbf{Xa}, \mathbf{Zb}). \tag{1}$$

Here ρ denotes the resulting canonical correlation. CCA gives a similarity spectrum between two representation spaces: higher canonical correlations indicate that the representations remain relatively stable across conditions, whereas lower correlations indicate stronger visually induced shifts. Implementation details are provided in Appendix A.7.

Interpretability. To examine how the models integrate textual and visual information when producing concreteness and imagery judgments, we conduct attribution analysis using Integrated Gradients (IG), implemented with Layer Integrated Gradients (LayerIG; Sundararajan et al., 2017) in Captum (Kokhlikyan et al., 2020). Let $F(\mathbf{x})$ denote the model output for input $\mathbf{x}$, and let $\mathbf{x}'$ denote a structure-preserving baseline input. For a token t, embedding dimension k, token embedding value $e_{t,k}$, and baseline value $e'_{t,k}$, IG is defined as

$$\begin{aligned}&\mathrm{IG}_{t,k}(\mathbf{x})\\ =&(e_{t,k} - e'_{t,k}) \int_0^1 \frac{\partial F\big(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\big)}{\partial e_{t,k}}\, d\alpha,\end{aligned} \tag{2}$$

where α interpolates along the path from $\mathbf{x}'$ to $\mathbf{x}$. Token-level attributions are then obtained by summing over embedding dimensions:

$$\mathrm{IG}_t = \sum_{k=1}^{d} \mathrm{IG}_{t,k}, \tag{3}$$

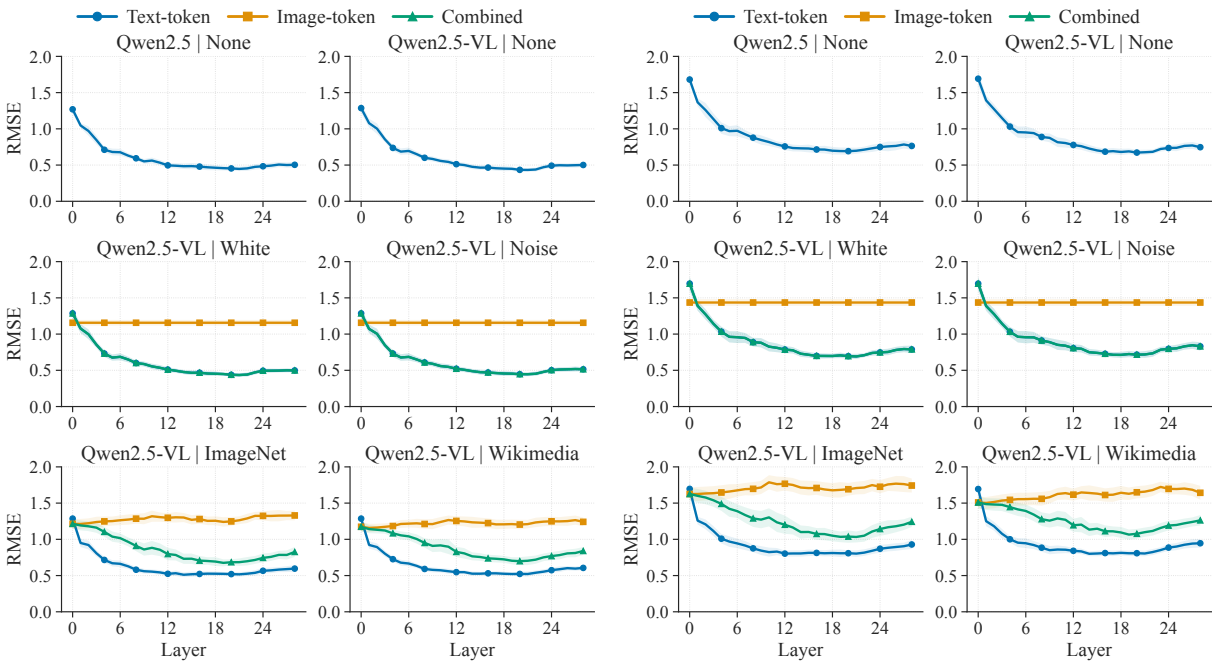
where d is the embedding dimension. We use a structure-preserving baseline and compute token-level attributions for both text and image tokens. This allows us to visualize which tokens drive a prediction and to summarize modality reliance by comparing the total attribution mass assigned to text tokens versus image tokens across visual conditions. The full formulation and implementation details are provided in Appendix A.8.

4 Results and Discussion

Our primary finding is that VLMs are vulnerable to spurious visual cues that can undermine their ability to capture human-like judgments of concreteness and imagery. When provided with real images, these models frequently shift toward stronger visual reliance, and this is often associated with worse performance than text-only models and VLMs with uninformative images, particularly on abstract/low-imagery items. The following sections present detailed analyses supporting this conclusion.

4.1 Model Predictions

Table 1 presents the RMSE of model predictions compared to human ratings for concreteness and imagery. Text-only models (LLaMA, Mistral, Qwen) are competitive with and often stronger than their vision–language counterparts in the *None* condition, indicating that vision fine-tuning does not inherently enhance performance on these tasks in our setup. Because the text-only model in each family serves as the backbone for its VLM counterpart, these comparisons are especially direct. Among VLMs, performance varies substantially based on the type of visual input provided. When real images are provided (from either ImageNet or Wikimedia), models often improve for concrete/high-imagery words but frequently worsen for abstract/low-imagery words. Overall



(a) MT40k (Concreteness)

(b) CP2004B (Imagery)

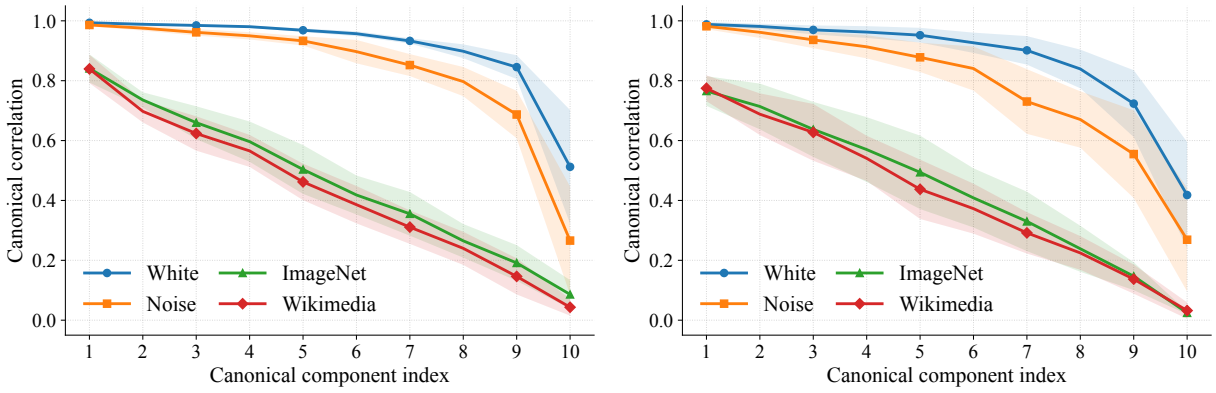
Figure 2: Layer-wise probing performance of Qwen2.5 and Qwen2.5-VL using ridge probes under different visual input conditions for (a) concreteness and (b) imagery. For Qwen2.5-VL, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

effects are therefore heterogeneous across models and subsets rather than uniformly positive or uniformly negative. The improved performance on concrete/high-imagery words suggests that real images can provide useful visual grounding for these concepts, which has been observed in prior work (Zhuang et al., 2024b,a). However, the substantial performance drop on abstract/low-imagery words is consistent with the hypothesis that real-image inputs can introduce spurious visual cues when image-word alignment is weak.

4.2 Probing

Figure 2 shows layer-wise probing performance for concreteness and imagery for Qwen2.5 and Qwen2.5-VL under different visual input conditions. Error bands indicate standard deviations across five prompt templates. For brevity, we present results using ridge regressors; results with other regressors and metrics exhibit similar trends (see Appendix B). For image inputs, we separately probe text-token and image-token hidden representations as well as the averaged hidden representations combining both modalities. For Qwen2.5, probing performance increases steadily across layers and stabilizes at higher layers, indicating that concreteness and imagery are robustly encoded in later hidden representations. For Qwen2.5-VL, a similar pattern is observed when no image is provided or when the input image is uninformative (white or random-noise), with probing curves closely matching those of text-only models.

However, conditioning on real images in our Qwen2.5-VL probing setup is associated with degraded probing performance across layers. The image-token hidden representations yield consistently poor performance, suggesting that visual features alone do not effectively capture concreteness or imagery information. The text-token hidden representations in the real-image condition also show reduced performance compared to text-only and uninformative conditions, indicating that real-image conditioning is associated with weaker encoding of these linguistic properties. When combining text and image hidden representations, performance remains in between that of text and image tokens alone, which is consistent with real-image inputs contributing visual signals that are not always helpful for this task. This pattern is consistent with prior findings of asymmetric cross-modal influence in multimodal transformers, where textual representations can be more strongly affected by visual



(a) MT40k (Concreteness)

(b) CP2004B (Imagery)

Figure 3: Canonical correlations between final-layer hidden representations obtained with different visual input conditions for Qwen2.5-VL on (a) concreteness and (b) imagery. The no-image visual input condition is compared against white images, random-noise images, and real images. Correlations are computed on the top 10 principal components obtained via PCA. Error bands indicate standard deviations across five prompt templates.

input than the reverse direction (Frank et al., 2021). In particular, although early and mid layers sometimes exhibit transient increases, performance in later layers, where semantic information is typically consolidated (Tenney et al., 2019; Hewitt and Liang, 2019), drops significantly relative to other conditions. Taken together, these results indicate that, in this setting, real-image conditioning can interfere with encoding of concreteness and imagery, likely when visual evidence is weakly aligned with the lexical target.

4.3 Canonical Correlation Analysis

Figure 3 presents the canonical correlations between the final-layer hidden representations obtained with different visual input conditions for Qwen2.5-VL. The no-image visual input condition is compared against white images, random-noise images, and real images. Hidden representations from uninformative images (white and random-noise) exhibit high canonical correlations with the no-image visual input condition across the top principal components, indicating that the model's hidden representations remain largely consistent regardless of the presence of uninformative visual input. In contrast, hidden representations derived from real images show markedly lower canonical correlations with the no-image visual input condition, with the last few components exhibiting correlations close to zero. This pattern suggests that real images are associated with significant changes in the model's hidden representations, reflecting a stronger influence of visual content. The pronounced divergence in canonical correlations indicates that real-image conditions are associated with representational shifts relative to the no-image condition. These findings align with the probing results and are consistent with visual influences that, for this task, can detract from the models' capacity to accurately capture concreteness and imagery.

4.4 Interpretability: A Case Study

Figure 4 presents a case study of Qwen2.5-VL on the target word *nature* (human concreteness rating: 2.92) using LayerIG. It demonstrate how the attribution landscape shifts between text and image tokens when the visual input changes from uninformative images to real images and how this shift is associated with changes in the model's prediction and alignment with human ratings.

With white and random-noise images, attributions are concentrated on text tokens, and the model predicts 3 in both cases (absolute error 0.08), which is close to the human ground truth. This indicates that when visual input is uninformative, the model relies primarily on linguistic cues and remains relatively well calibrated.

However, when conditioned on real images (ImageNet and Wikimedia), the attribution landscape shifts. The real images depict natural outdoor scenes that are visually rich and vivid, even though they may not be semantically aligned with the abstract concept of *nature* as rated by humans. As a result, image tokens receive substantially higher attribution mass, and the model prediction shifts upward to 5 in both cases (absolute error 2.08). This overestimation is consistent with our broader quantitative analyses suggesting that real-image condi-

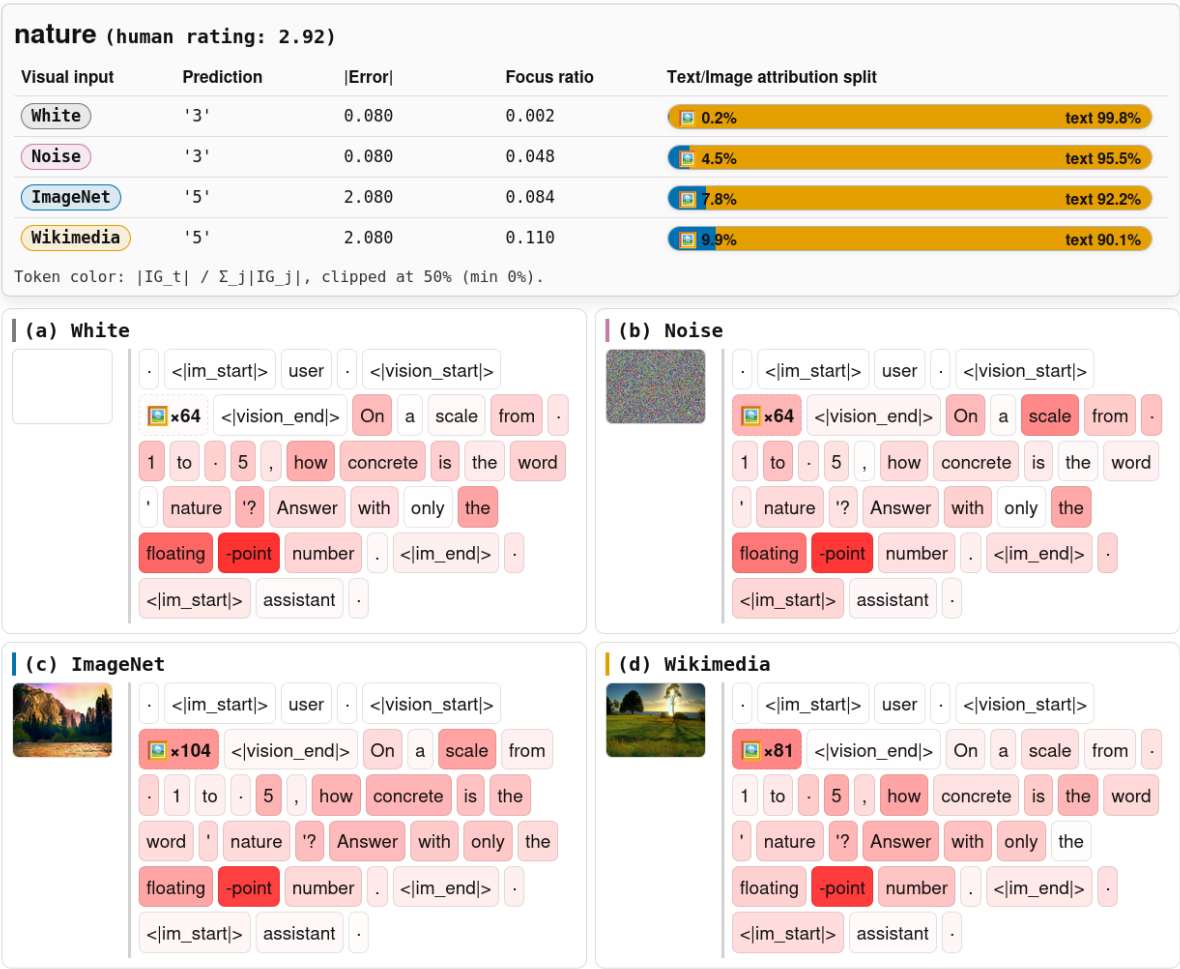


Figure 4: Integrated Gradients visualization for Qwen2.5-VL on the target word *"nature"* (human rating: 2.92) under four visual input conditions: White, Noise, ImageNet, and Wikimedia. The header summarizes the model's predicted next token, absolute error with respect to the human rating, focus ratio (image/text attribution mass), and modality-level attribution split. Token color indicates normalized attribution magnitude ($|IG_t|/\sum_j |IG_j|$), and image-token attributions are collapsed into a single block for readability.

tioning can bias the model toward higher concreteness judgments when visual evidence is weakly aligned with the lexical target.

5 Mitigation Strategies

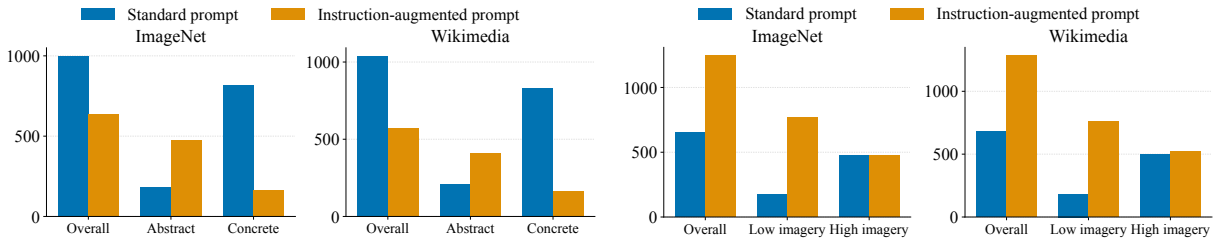
Our preceding analyses showed that real-image inputs can shift hidden representations, increase attribution to image tokens, and worsen alignment with human ratings, especially when the images are only weakly related to the target word. One plausible interpretation is that the model exhibits a form of shortcut learning at inference time: once a real-image input is present, it may rely on salient visual content as an easy cue, even when that cue is not reliably informative for the lexical property being rated (Geirhos et al., 2020). If this behavior is partly driven by prompt interpretation rather than by model weights alone, then making the possible irrelevance of the image explicit may help the model discount misleading visual evidence.

We therefore test a simple prompting intervention, using Qwen2.5-VL as the representative model. We prepend the following instruction to the original rating prompt:

> *The image may be irrelevant to the word; please base your rating solely on the word itself.*

This intervention, in line with the instructed prompting idea by Shi et al. (2023), does not alter the model's parameters or representations directly; instead, it changes the decision context by making the possible irrelevance of the image explicit.

Figure 5 compares the standard and instruction-augmented prompts by counting how often each



(a) MT40k (Concreteness)

(b) CP2004B (Imagery)

Figure 5: Effect of instruction-augmented prompting on Qwen2.5-VL performance in real-image visual input conditions for (a) concreteness and (b) imagery. We compare predictions from the standard prompt and the instruction-augmented prompt against human ratings, counting how often each variant yields a smaller absolute error (ties removed). Results are shown overall and separately for abstract (concreteness < 3) and concrete (concreteness >= 3) words in MT40k, and for low-imagery (imagery < 4) and high-imagery (imagery >= 4) words in CP2004B.

produces the smaller absolute error against human ratings under real-image conditions. The effect is selective rather than uniform. On MT40k, the instruction-augmented prompt is preferred for abstract words, but the standard prompt remains stronger overall and for concrete words. On CP2004B, the instruction-augmented prompt is preferred overall and especially for low-imagery words, whereas the high-imagery subset is close to evenly split. These results indicate that the intervention can mitigate some of the harm from real-image conditioning, but mainly for the subsets that are most vulnerable to visually induced overestimation, rather than as a universal improvement.

Because this intervention operates entirely at inference time, it can be applied without architectural changes, additional supervision, or retraining. More broadly, these results suggest that prompt design can function as a lightweight guardrail when visual inputs are noisy, weakly aligned, or incidental, although stronger and more consistent mitigation will likely require methods that adaptively calibrate the contribution of visual evidence rather than relying on a single fixed instruction.

6 Conclusion

In this work, we investigate the limitations of modern instruction-tuned VLMs in capturing human-like judgments of word concreteness and imagery. Our comprehensive evaluation shows that these models' visual grounding capabilities can be undermined by their vulnerability to spurious visual cues. Through probing, canonical correlation analysis, and interpretability techniques, we show that VLMs can over-rely on visual information even when it is irrelevant or misleading, and this is associated with reduced linguistic fidelity in our evaluation setting. To address this issue, we propose a simple mitigation strategy: augmenting prompts with explicit instructions to focus solely on textual content. In our experiments, this approach yields pronounced gains on abstract and low-imagery words that appear most vulnerable to visually induced errors, highlighting the potential of inference-time interventions to counteract such biases in visual grounding. Overall, our findings underscore the importance of carefully considering the role of visual inputs in multimodal models and pave the way for future research aimed at enhancing their alignment with human cognition.

Limitations

Our study has three main limitations. First, evaluation is limited to English lexical norms (MT40k and CP2004B) and single-word items, so findings may not transfer directly to other languages, multi-word expressions, or sentence-level grounding. Second, our real-image conditions rely on top-k retrieval from ImageNet and Wikimedia without full manual relevance annotation, so we cannot always separate retrieval mismatch from the model's own visual over-reliance, especially for abstract and low-imagery words. Third, our mitigation study is narrow: we test one inference-time instruction on Qwen2.5-VL, and the gains are selective rather than universal. The prompt helps most on subsets that are especially vulnerable to visually induced overestimation, but it may be insufficient when visual evidence is partly relevant. Different model families, prompts, or datasets may therefore lead to different conclusions.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Singh Chaplot, Jessica Chudnovsky, Saurabh Garg, Théophile Gervet, Soham Ghosh, Am'elie H'eliou, Paul Jacob, Albert Q. Jiang, Timothée Lacroix, Guillaume Lample, Diego de Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, and 18 others. 2024. Pixtral 12b. *ArXiv preprint*, abs/2410.07073.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is BERT blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6778–6788. IEEE.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *ArXiv preprint*, abs/2502.13923.

Kai Barthel, Nico Hezel, Konstantin Schall, and Klaus Jung. 2023. navigu.net: Navigation in visual image graphs gets user-friendly. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

James M. Clark and Allan Paivio. 2004. Extensions of the paivio, yuille, and madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36:371–383.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Friendly, Patricia E. Franklin, David Hoffman, and David C. Rubin. 1982. The toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14:375–399.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265, Doha, Qatar. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard

Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *ArXiv preprint*, abs/2009.07896.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. 2023. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, Toronto, Canada. Association for Computational Linguistics.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76 1:Suppl:1–25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv preprint*, abs/2412.15115.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian,

Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, and 29 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv preprint*, abs/2504.10479.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024a. Lexicon-level contrastive visual-grounding improves language modeling. *ArXiv preprint*, abs/2403.14551.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024b. Visual grounding helps learn word meanings in low-data regimes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1311–1329, Mexico City, Mexico. Association for Computational Linguistics.

A Implementation Details

A.1 Model Checkpoints

Table 2 lists the model identifiers used in our pipeline. All models are loaded via their Hugging Face implementations.

Model family	Hugging Face identifier
LLaMA 2	`meta-llama/Llama-2-7b-chat-hf`
Mistral	`mistralai/Mistral-7B-Instruct-v0.3`
Qwen2.5	`Qwen/Qwen2.5-7B-Instruct`
LLaVA 1.5	`llava-hf/llava-1.5-7b-hf`
InternVL3	`OpenGVLab/InternVL3-8B-hf`
Pixtral	`mistral-community/pixtral-12b`
Qwen2.5-VL	`Qwen/Qwen2.5-VL-7B-Instruct`
Gemma 3	`google/gemma-3-4b-it`

Table 2: Model checkpoints used in the experiments.

A.2 Evaluation Set Construction

We construct the evaluation set in three steps. First, from MT40k we retain only valid single-word entries: we drop missing ratings, remove bigrams and hyphenated forms, lowercase and trim surface forms, and deduplicate by word. Second, from CP2004B we retain valid single-word entries: we drop missing values and non-positive imagery scores, remove hyphenated forms, lowercase and trim tokens, and deduplicate by word. Third, for direct comparability across concreteness and imagery, we evaluate on the shared vocabulary intersection between these processed sets. This yields a shared evaluation set used across both tasks.

A.3 Prompt Variants

For each word, we query each model with five prompt templates and average parsed numeric predictions across the five responses. Below are the templates used for concreteness rating:

1. *On a scale from 1 to 5, how concrete is the word '{word}'? Answer with only the floating-point number.*
2. *Rate the concreteness of the word '{word}' from 1 (very abstract) to 5 (very concrete). Answer with only the floating-point number.*
3. *How tangible or concrete is the concept of '{word}'? Give a score from 1 to 5. Answer with only the floating-point number.*
4. *Evaluate how physically grounded the word '{word}' is, using a scale from 1 to 5. Answer with only the floating-point number.*
5. *Is the word '{word}' something you can perceive with your senses? Rate its concreteness from 1 to 5. Answer with only the floating-point number.*

The templates used for imagery rating are:

1. *On a scale from 1 to 7, how easy is it to form a mental image of the word '{word}'? Answer with only the floating-point number.*
2. *How easy is it to picture the word '{word}' in your mind, from 1 (hard) to 7 (easy)? Answer with only the floating-point number.*
3. *Rate the imagery of the word '{word}' on a 1–7 scale. Answer with only the floating-point number.*
4. *Imagine the word '{word}'. How vivid is the mental image? Rate from 1 to 7. Answer with only the floating-point number.*
5. *How imageable is the word '{word}'? Respond with a number from 1 (least) to 7 (most). Answer with only the floating-point number.*

A.4 Visual Input Conditions

We evaluate the same five visual input conditions used in the main paper: *None*, *White*, *Noise*, *ImageNet*, and *Wikimedia*. In the pipeline, the *Noise* condition is implemented as a random-noise placeholder image. For real-image settings, each word is associated with a directory of up to five retrieved images, and image assignment is shuffled before prompt-template pairing.

A.5 Output Parsing and Aggregation

Model outputs are obtained by greedy decoding with a maximum generation length of 128 tokens. To extract the predicted numeric rating from the model output, we identify all numeric spans in the generated text and keep the last value within the valid task range. Predictions outside the valid task range are discarded (1–5 for concreteness, 1–7 for imagery). For each word, the final prediction is the mean of valid parsed numbers across the five prompt variants. If no valid number is found, the item is treated as missing.

A.6 Probing Details

Probing uses 10-fold cross-validation on hidden representations. We report probing results with three regressor architectures: ridge, SVM, and

MLP. For ridge, we use scikit-learn's `Ridge` model. For SVM, we use scikit-learn's `SVR` model with an RBF kernel. For MLP, we use a one-hidden-layer regressor (hidden size 2048, ReLU, dropout 0.2) trained with Adam for 100 epochs (batch size 512, learning rate 0.001, weight decay 1×10^{-5}). In all cases, features are scaled with min-max normalization. All runs use random seed 42.

A.7 CCA Details

For each word, we extract final-layer hidden representations under each visual input condition and align words across the two conditions being compared (e.g., *None* vs. White/Noise/real-image conditions). This yields paired representation matrices $\mathbf{X}$ and $\mathbf{Z}$, where rows correspond to words and columns to hidden dimensions. More concretely, if $\mathbf{x}_i, \mathbf{z}_i \in \mathbb{R}^d$ denote the paired final-layer hidden representations for word i, then $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$.

CCA then finds linear projection vectors $\mathbf{a}$ and $\mathbf{b}$ that maximize the correlation between the projected representations:

$$\rho = \max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^\top \Sigma_{XZ} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{XX} \mathbf{a}\, \mathbf{b}^\top \Sigma_{ZZ} \mathbf{b}}}, \tag{4}$$

Here ρ denotes the resulting canonical correlation, and Σ_{XX}, Σ_{ZZ}, and Σ_{XZ} are the empirical covariance and cross-covariance matrices. Repeating this procedure yields a spectrum of canonical correlations $\{\rho_k\}_{k=1}^{K}$, which summarizes the similarity between the two representation spaces.

Before running CCA, we standardize each matrix feature-wise across words. To improve numerical stability and reduce computational cost, we then apply Principal Component Analysis (PCA) separately to each matrix and retain the top $K = 10$ components. CCA is performed on these reduced matrices to compute component-wise canonical correlations. In Figure 3, each curve reports the mean canonical correlation across prompt variants, and the error band denotes one standard deviation across prompt variants.

A.8 Interpretability Details

For attribution analysis, we use Captum's Layer Integrated Gradients (LayerIG) on the first language transformer block. The attribution target is the logit of the model's predicted next token (argmax at the final position). In our runs, we use $n_\text{steps} = 8$ and prompt variant index 0.

The baseline is structure-preserving: template and special tokens remain unchanged, the attention mask is fixed, the user text span is replaced with a padding token (or EOS if padding is unavailable), and visual input is replaced with a black-image baseline in normalized pixel space.

Let $F(\mathbf{x})$ denote the model output for input $\mathbf{x} = (\mathbf{x}^{\text{text}}, \mathbf{x}^{\text{img}})$, and let $\mathbf{x}'$ denote the corresponding structure-preserving baseline input. For a token t, let $\mathbf{e}_t \in \mathbb{R}^d$ be its embedding vector and $\mathbf{e}'_t \in \mathbb{R}^d$ the corresponding baseline embedding, with $e_{t,k}$ and $e'_{t,k}$ denoting their k-th dimensions. The dimension-wise Integrated Gradients are

$$\mathrm{IG}_{t,k}(\mathbf{x}) = (e_{t,k} - e'_{t,k}) \int_0^1 \frac{\partial F\big(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\big)}{\partial e_{t,k}}\, d\alpha. \tag{5}$$

Here α interpolates along the path from $\mathbf{x}'$ to $\mathbf{x}$. Token-level attributions are then obtained by summing signed attributions across embedding dimensions:

$$\mathrm{IG}_t = \sum_{k=1}^{d} \mathrm{IG}_{t,k}. \tag{6}$$

In practice, the path integral is approximated using a finite number of interpolation steps. This formulation satisfies the completeness property, meaning that the sum of token attributions equals the difference between the model outputs at $\mathbf{x}$ and $\mathbf{x}'$.

For our case study, we convert these signed token attributions into absolute attribution mass. Specifically, we report normalized token intensity $|IG_t| / \sum_j |IG_j|$ and compute modality-level attribution splits as

$$A_{\text{modality}} = \sum_{t \in \mathcal{T}_{\text{modality}}} |\mathrm{IG}_t|, \tag{7}$$

for text tokens and image tokens in Figure 4.

A.9 Hardware

Experiments were run on NVIDIA L40S and H100 GPUs (48GB and 80GB memory, respectively). Model loading uses Hugging Face `device_map="auto"`, so computation is placed automatically on available GPU device(s).

B Additional Probing Results

Model prediction results are reported in the main paper. This appendix section reports supplementary probing results for additional models not shown in Figure 2, along with permuted-label baselines.

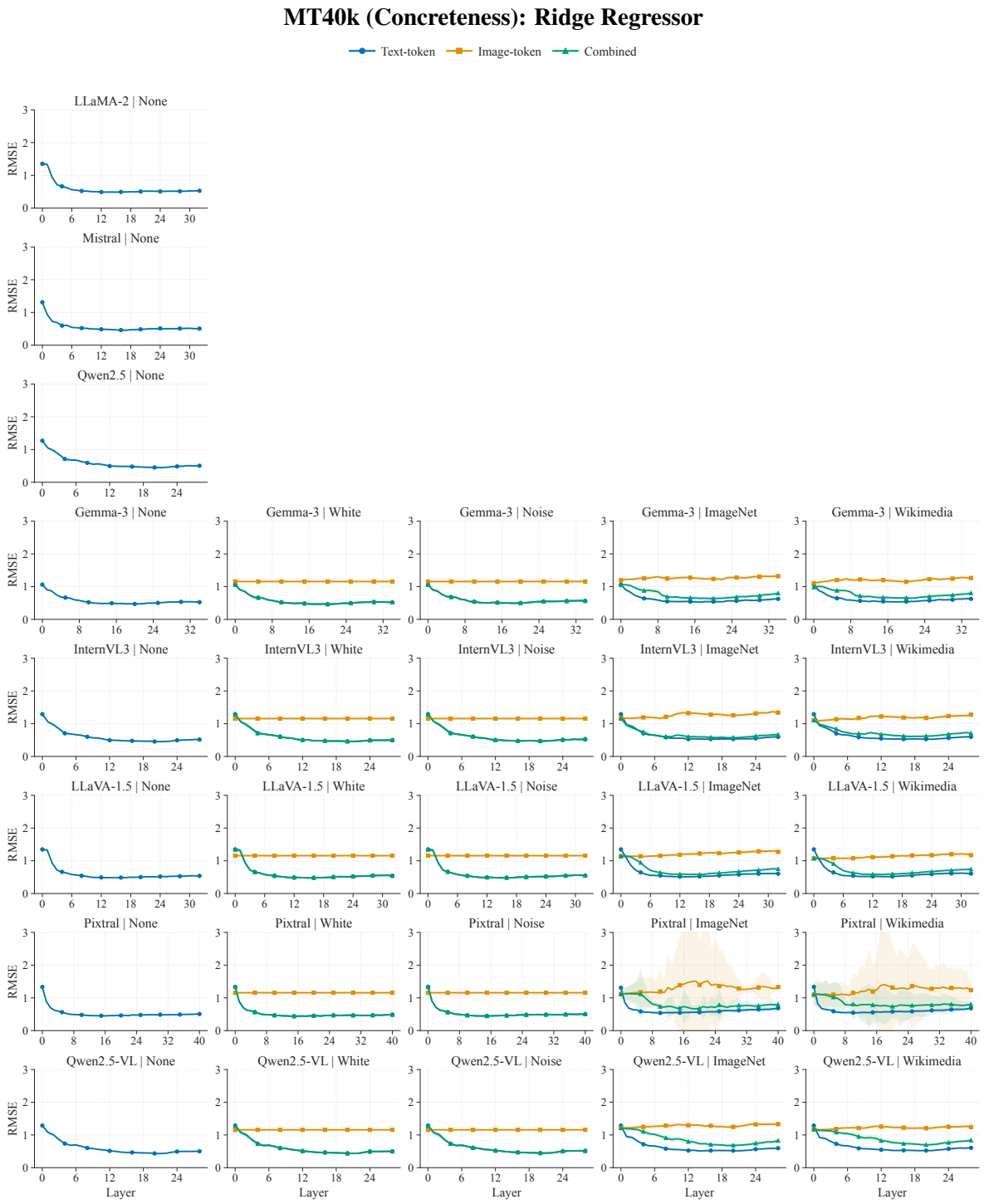


Figure 6: Supplementary layer-wise probing performance on MT40k (concreteness) using a ridge regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

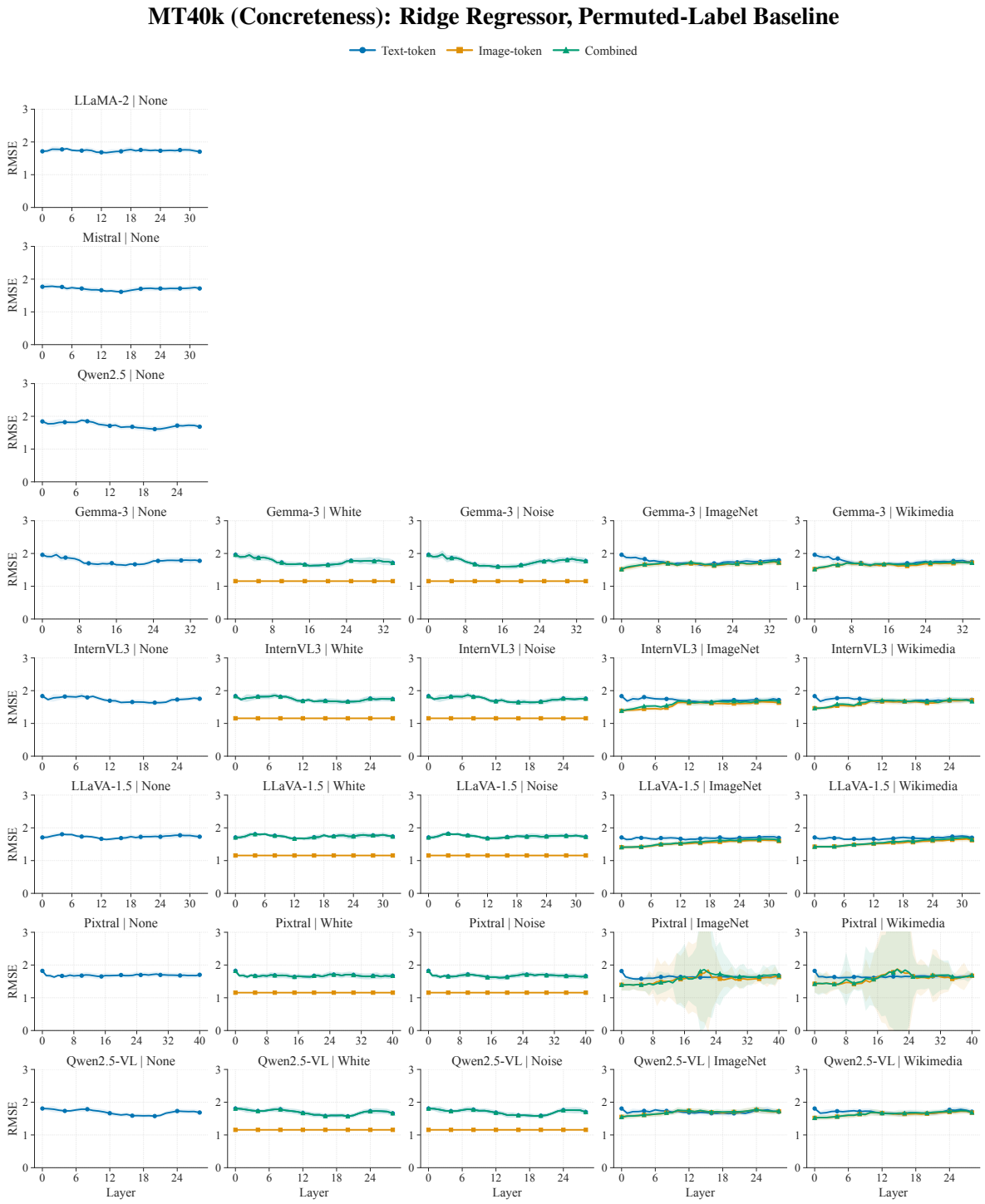


Figure 7: Permuted-label baseline for layer-wise probing performance on MT40k (concreteness) using a ridge regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

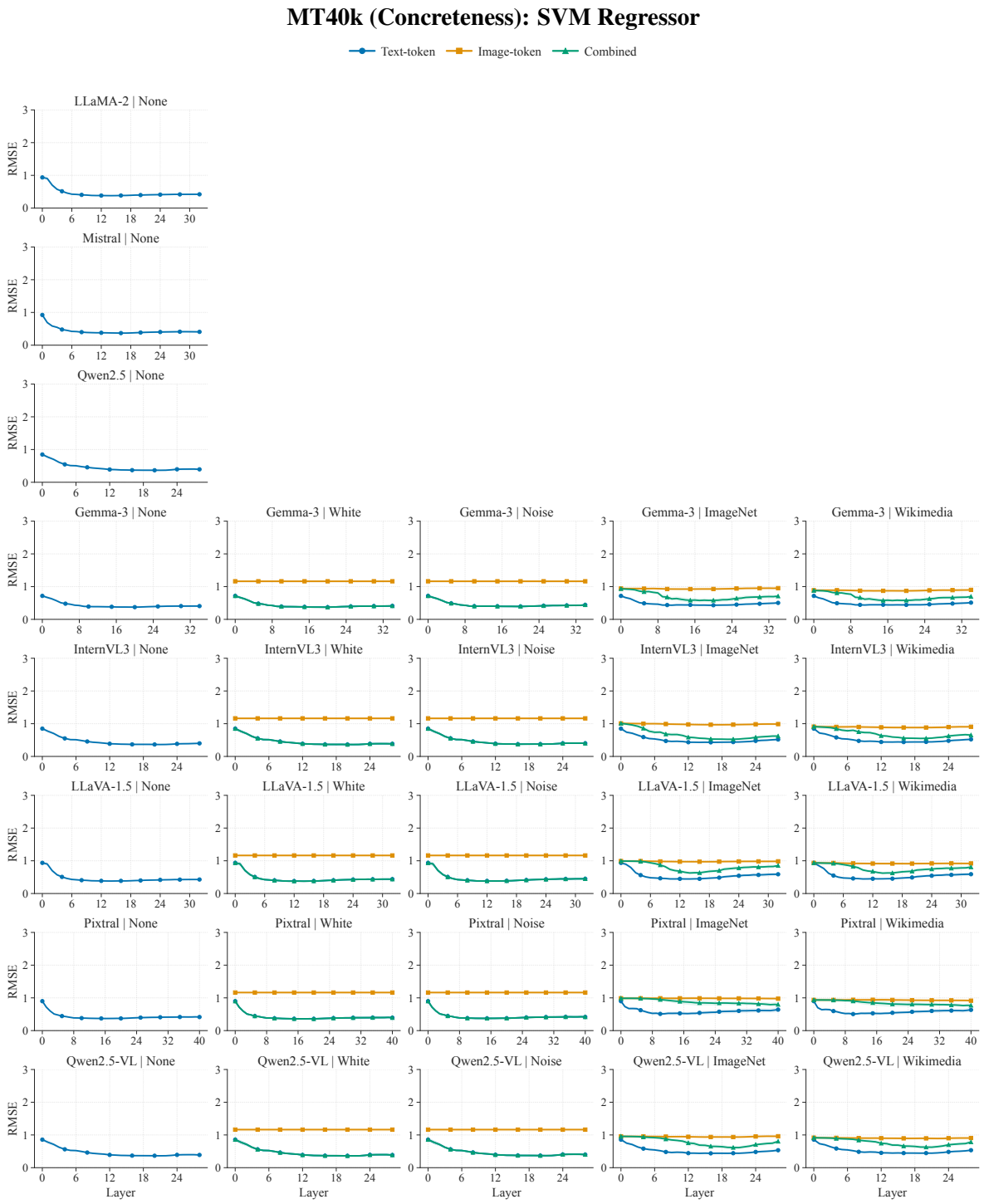


Figure 8: Supplementary layer-wise probing performance on MT40k (concreteness) using an SVM regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

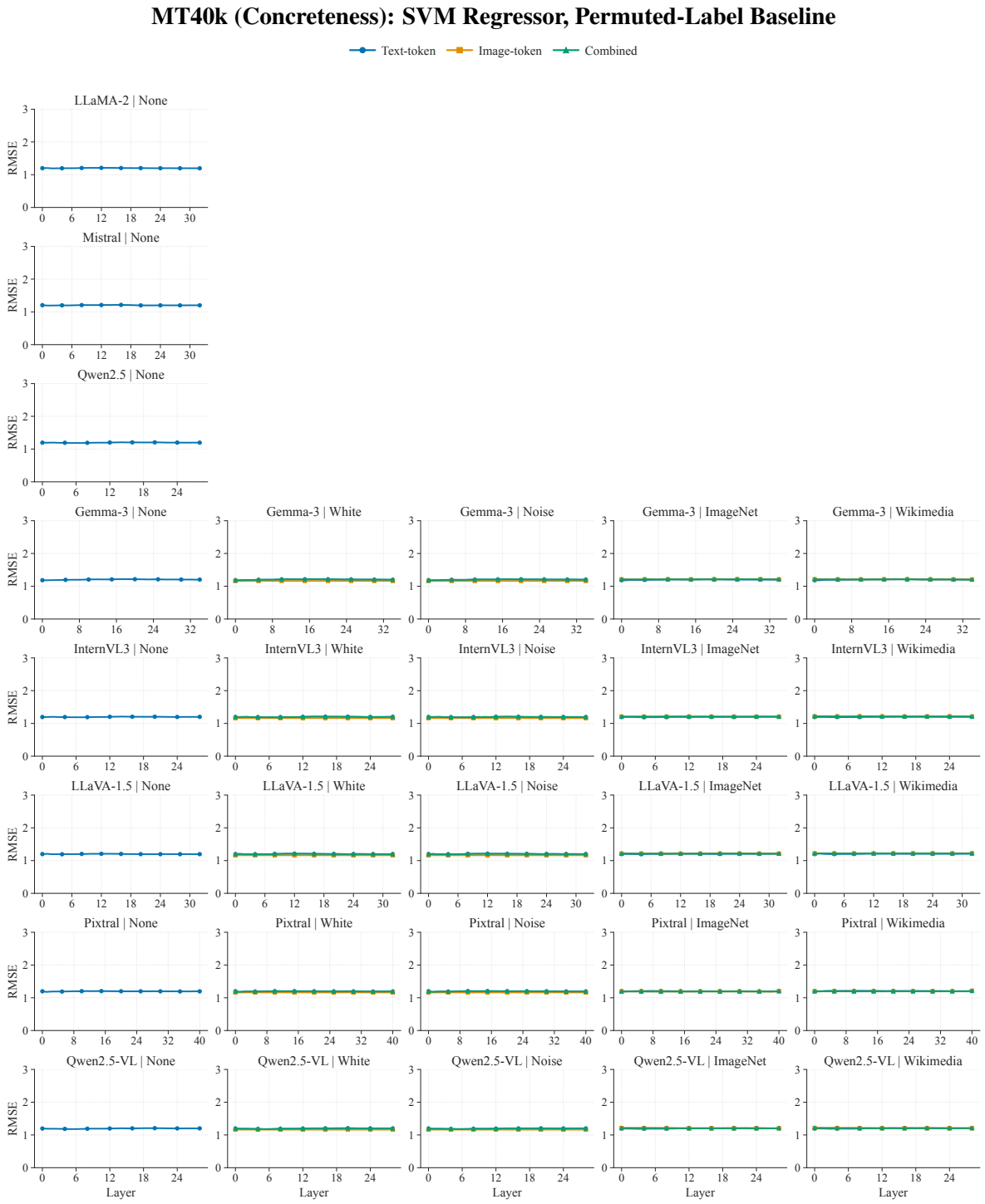


Figure 9: Permuted-label baseline for layer-wise probing performance on MT40k (concreteness) using an SVM regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

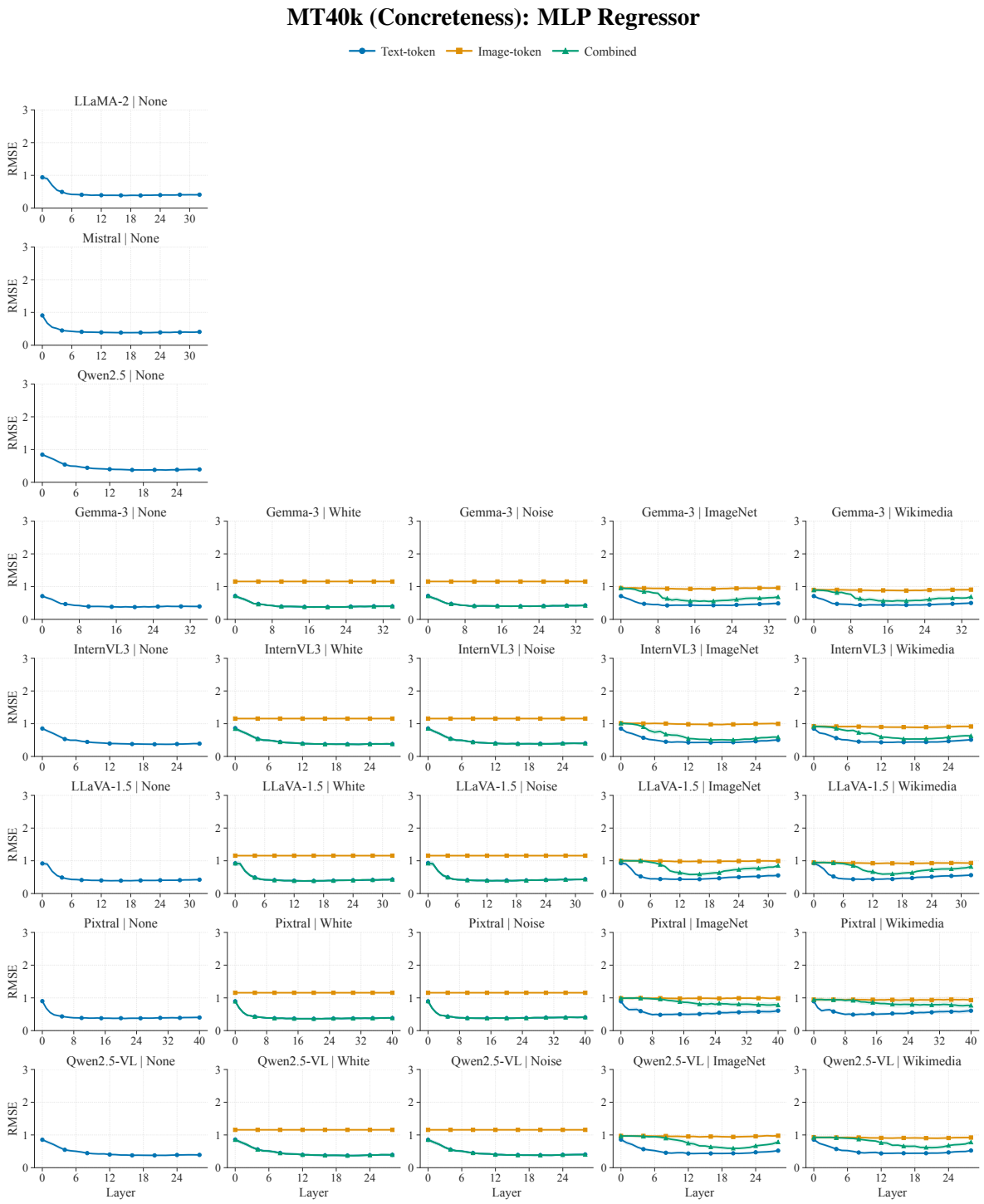


Figure 10: Supplementary layer-wise probing performance on MT40k (concreteness) using an MLP regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

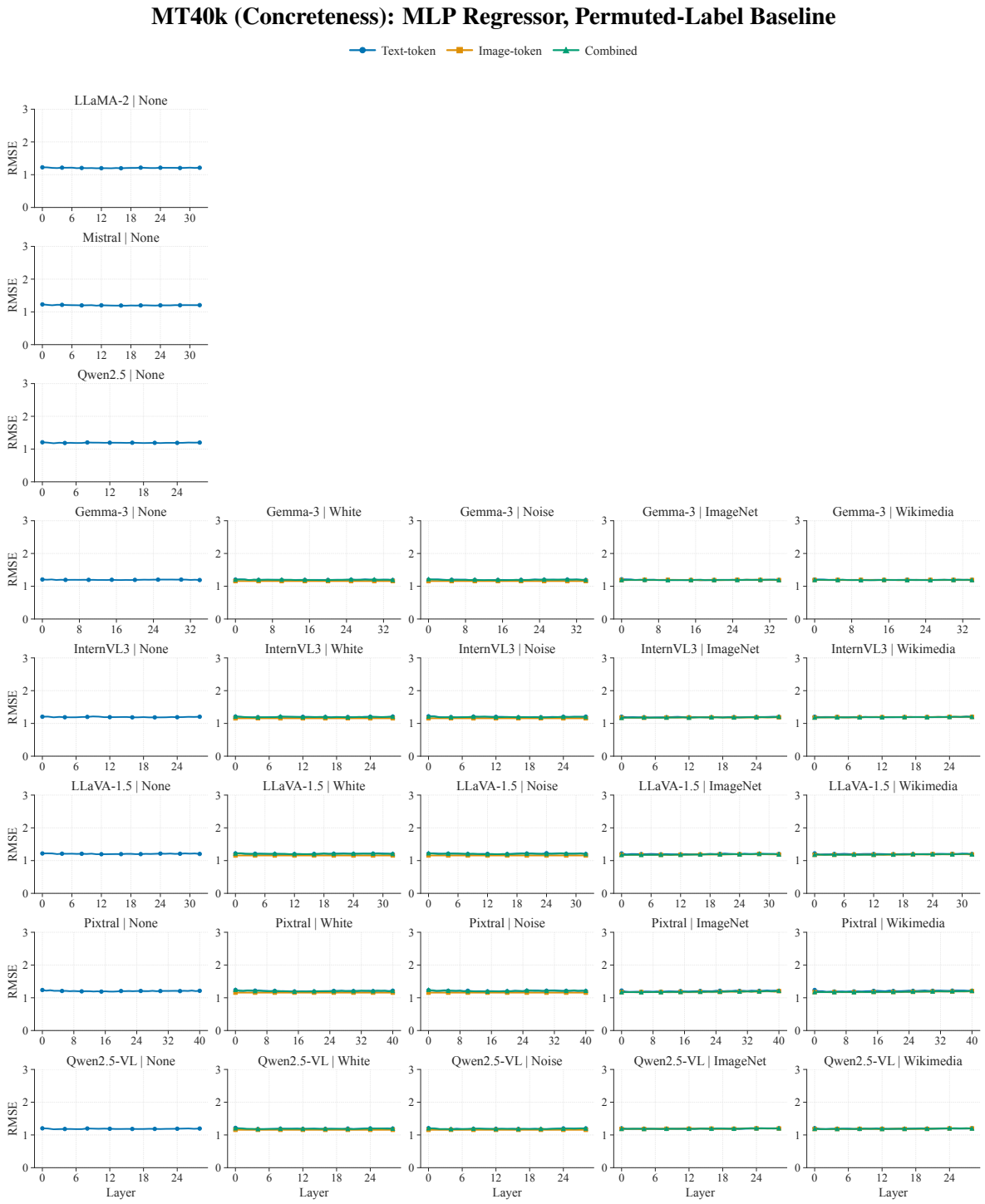


Figure 11: Permuted-label baseline for layer-wise probing performance on MT40k (concreteness) using an MLP regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

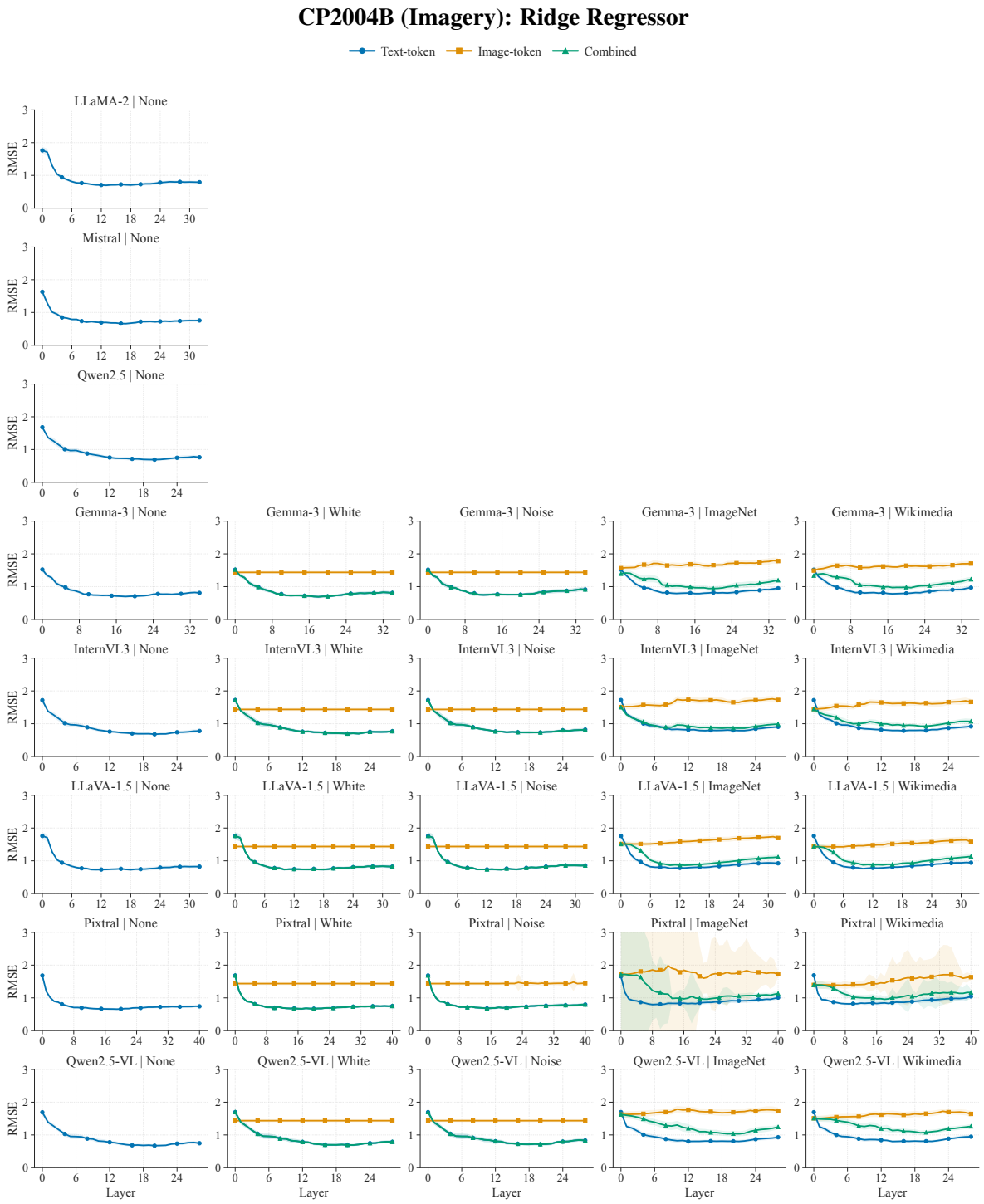


Figure 12: Supplementary layer-wise probing performance on CP2004B (imagery) using a ridge regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

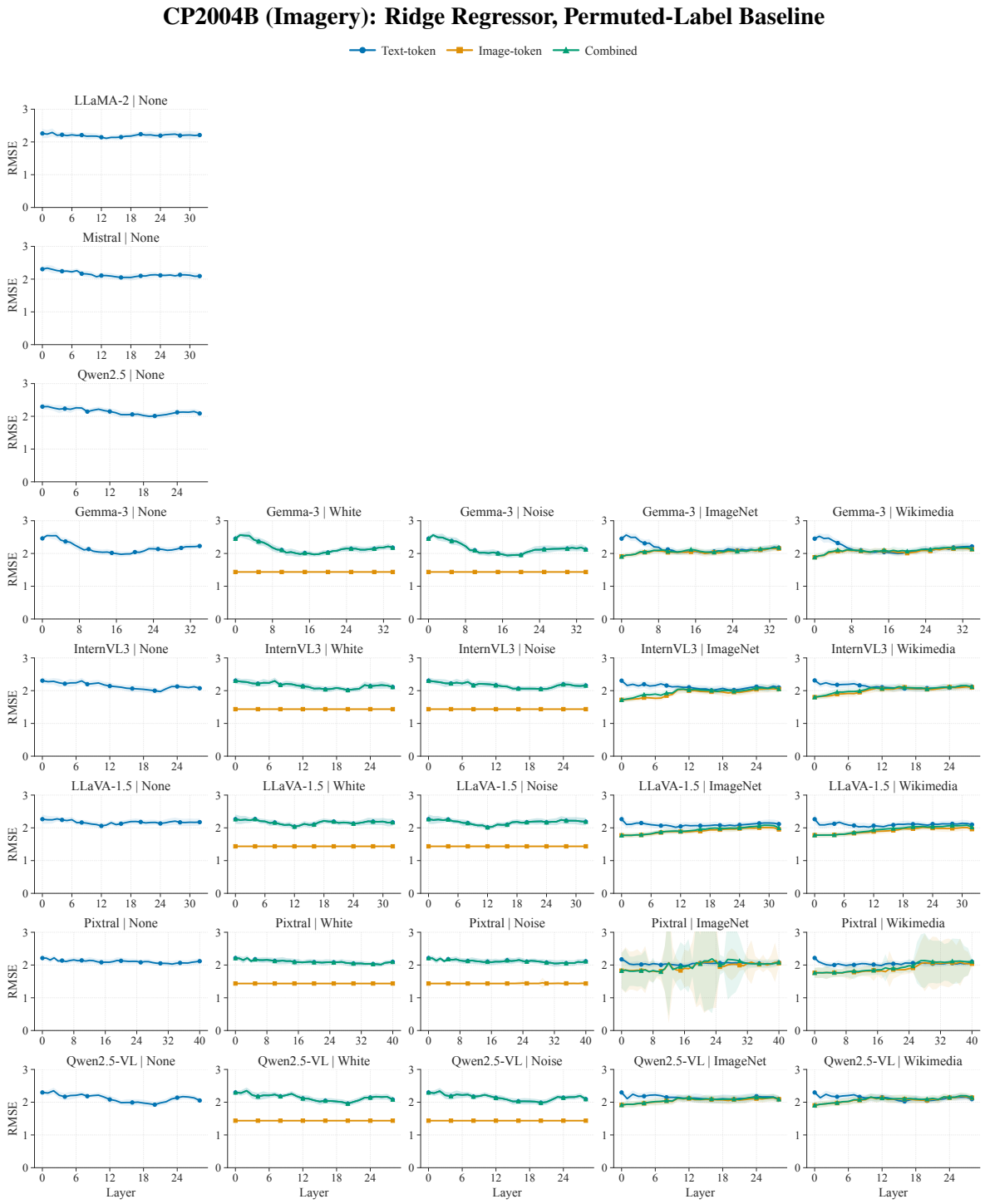


Figure 13: Permuted-label baseline for layer-wise probing performance on CP2004B (imagery) using a ridge regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

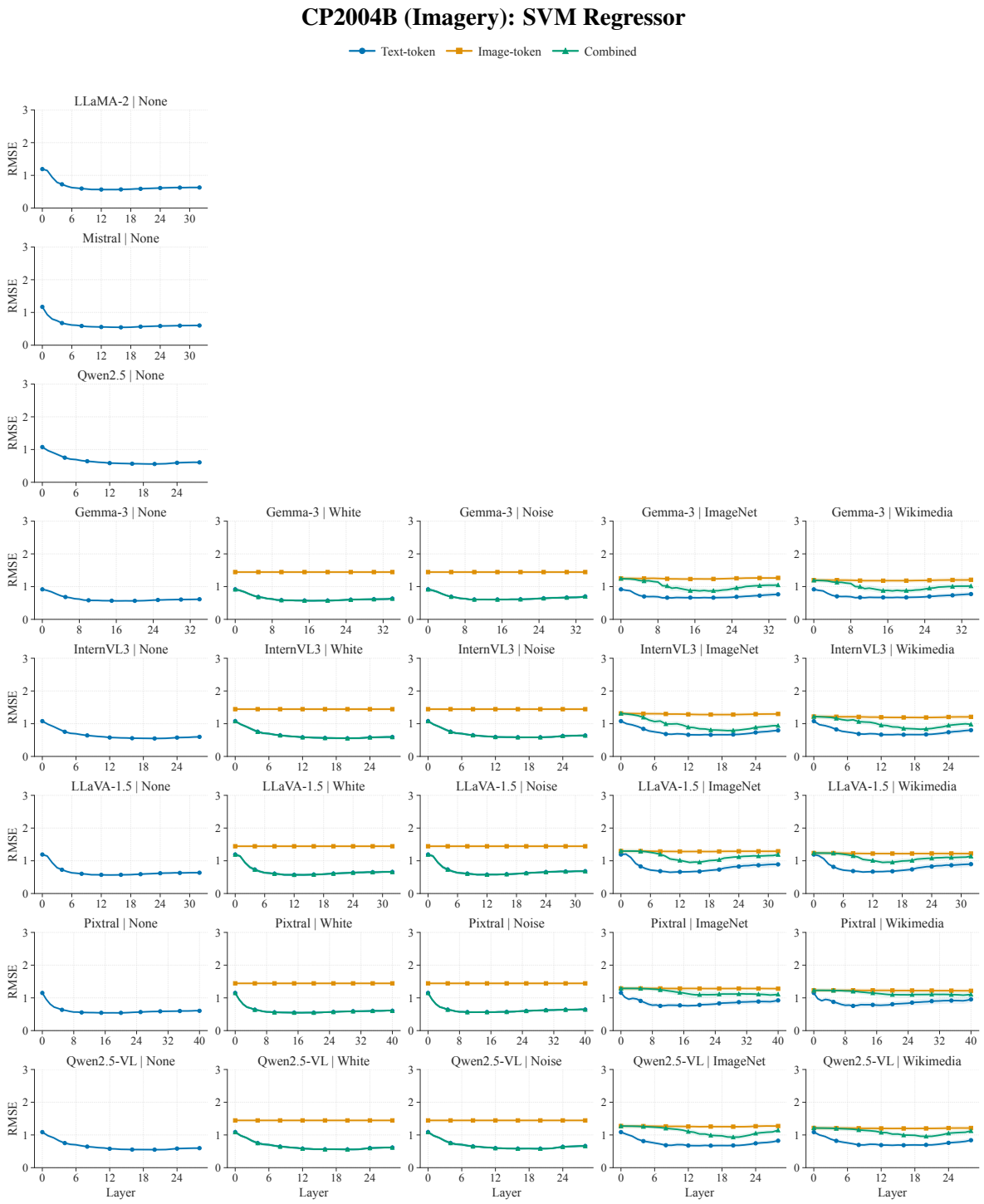


Figure 14: Supplementary layer-wise probing performance on CP2004B (imagery) using an SVM regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

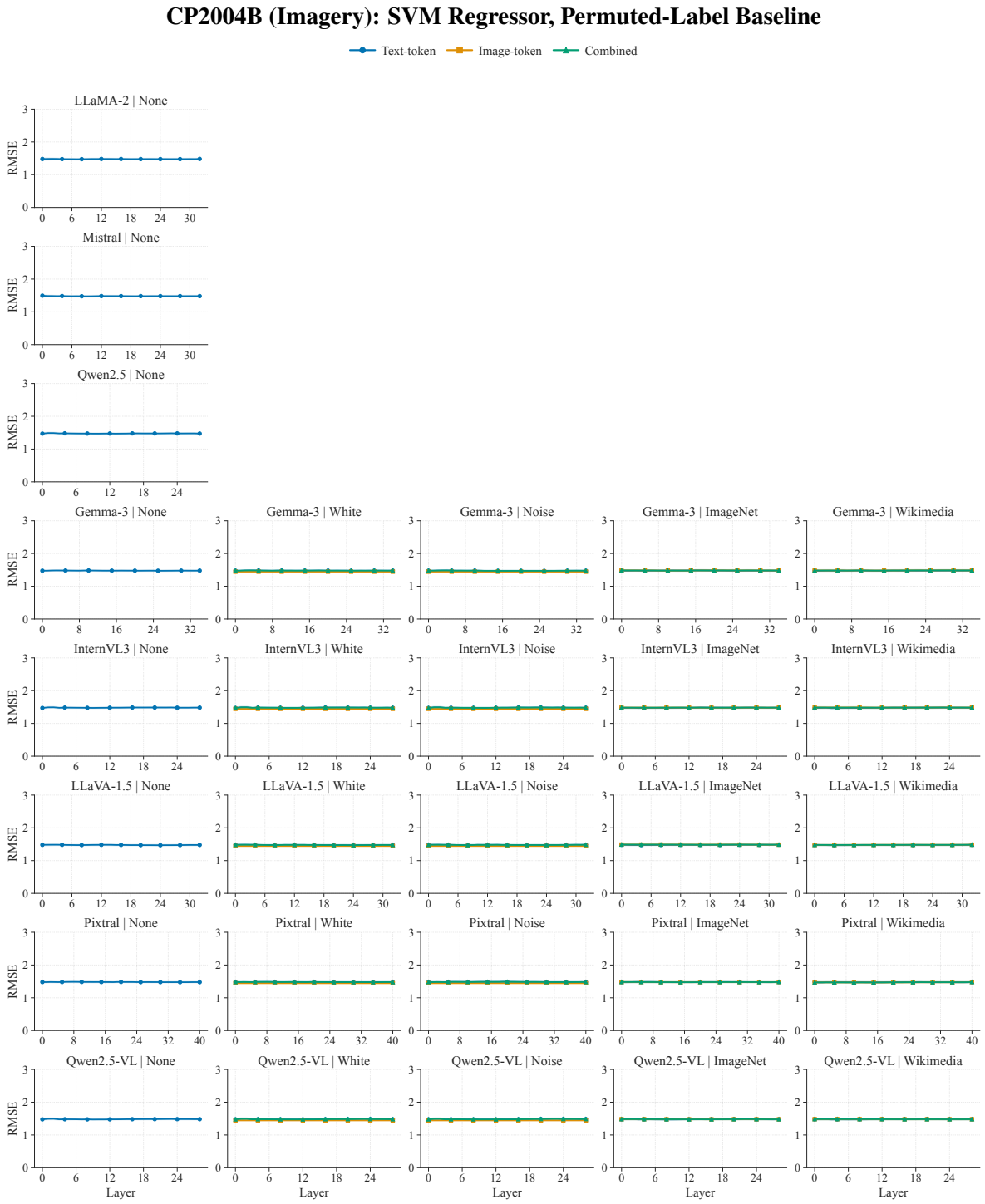


Figure 15: Permuted-label baseline for layer-wise probing performance on CP2004B (imagery) using an SVM regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

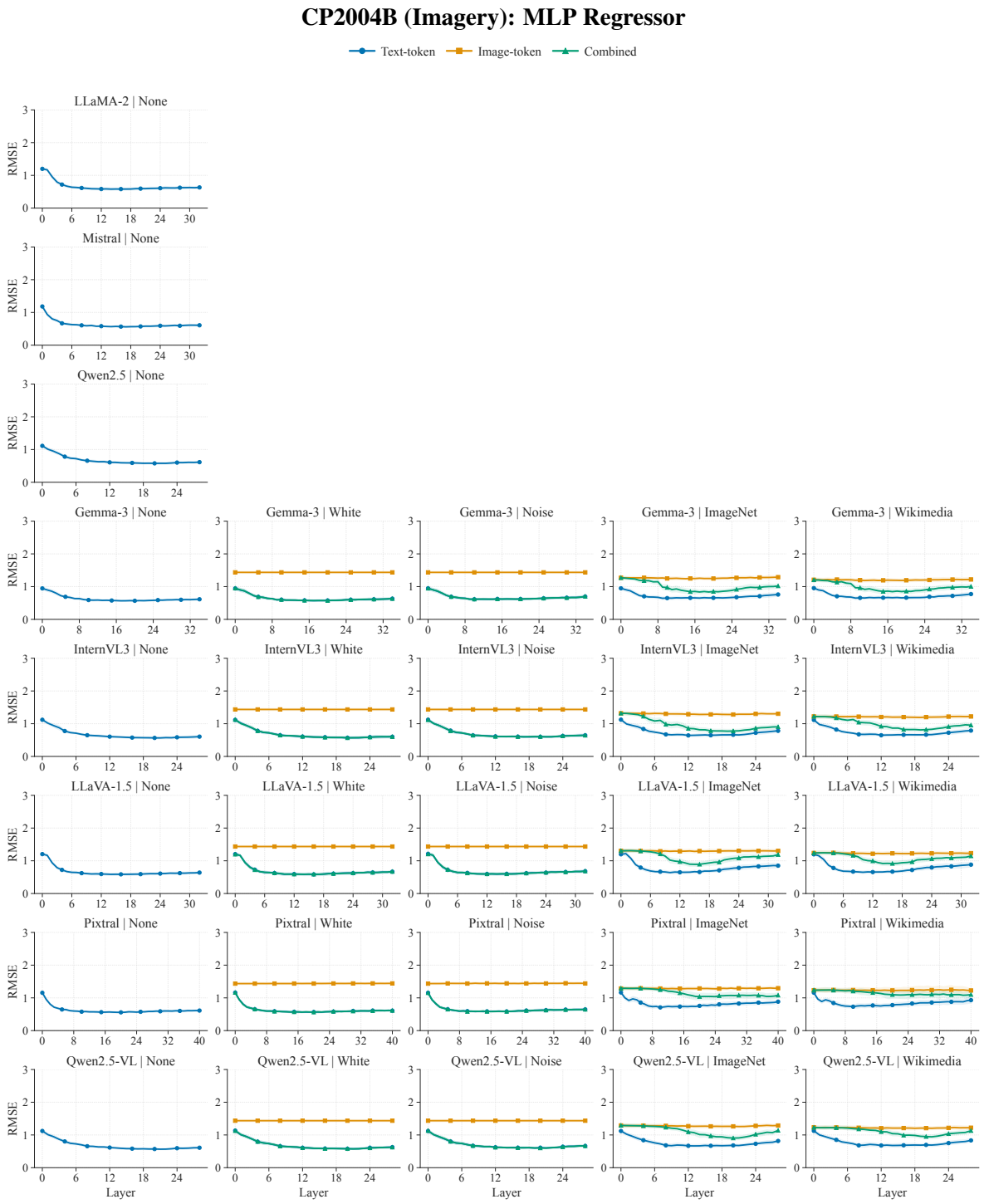


Figure 16: Supplementary layer-wise probing performance on CP2004B (imagery) using an MLP regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.

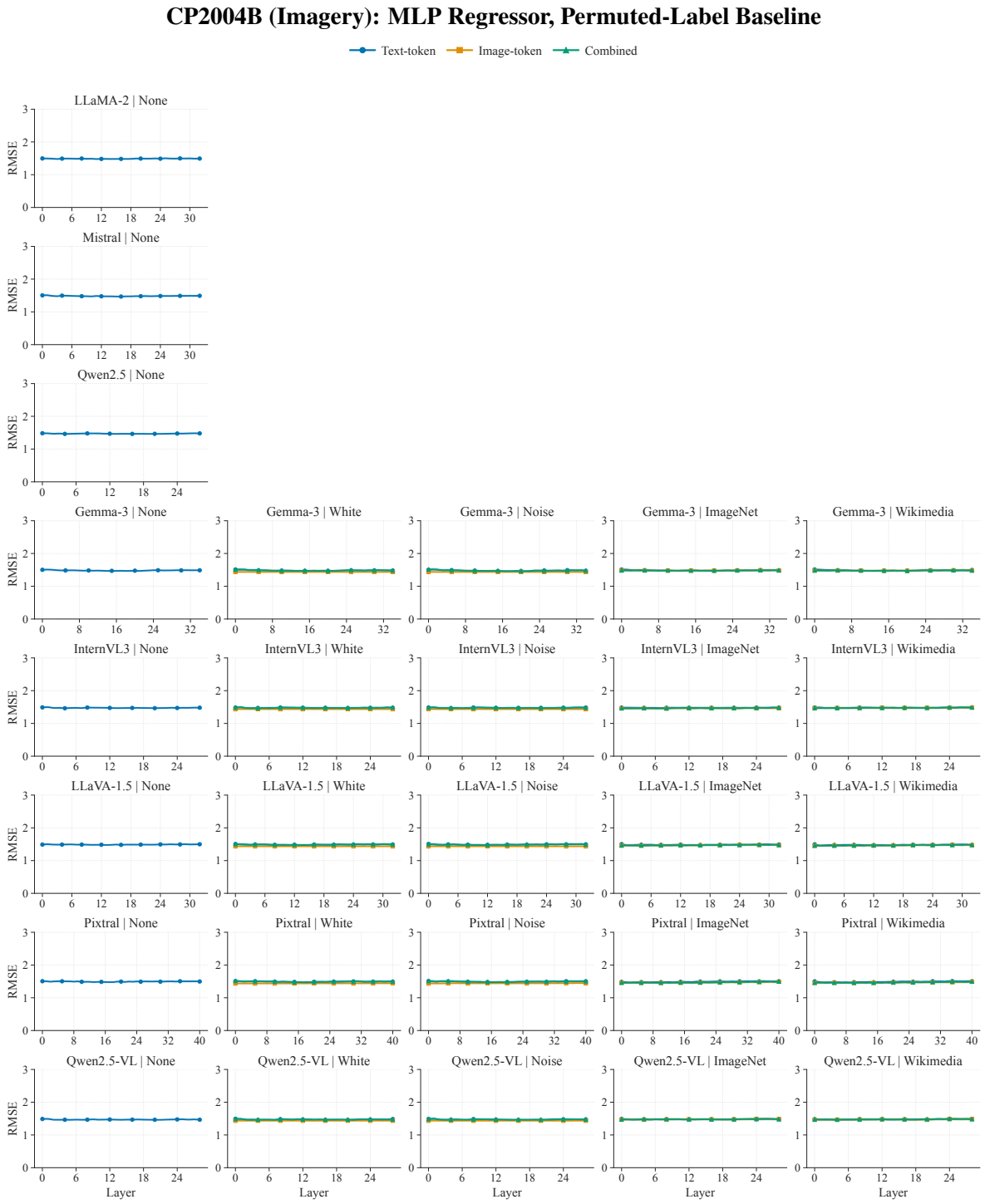


Figure 17: Permuted-label baseline for layer-wise probing performance on CP2004B (imagery) using an MLP regressor. For VLMs, probing is performed separately on text-token hidden representations, image-token hidden representations, and their combined hidden representations. Error bands indicate standard deviations across five prompt templates.