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

**Anonymous ACL submission**

## Abstract

Visual inputs are often assumed to improve language understanding in multimodal models. We examine this assumption by asking whether vision–language models (VLMs) can distinguish useful visual evidence from incidental image context in lexical judgments. We use human concreteness and imagery ratings because they span words with varying expected visual relevance, from abstract and low-imagery words to concrete and high-imagery words. We find that real-image contexts do not yield consistent gains and often hurt alignment with human ratings, most sharply when visual evidence is least relevant. Through probing and canonical correlation analysis, complemented by an attribution case study, we find that real-image contexts are associated with representational shifts and greater sensitivity to spurious visual cues, coinciding with weaker recoverability of the targeted lexical properties. We further show that instructing models to focus solely on textual content at inference time can reduce this degradation, with the clearest gains on these vulnerable subsets. Our findings suggest that current instruction-tuned VLMs need better calibration of when visual context should inform lexical judgments.

## 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 using visual context in ways that align with human interpretation. Prior work shows that models can over-rely on superficial visual–textual correlations (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 when models use visual context as meaningful evidence rather than as a shortcut, especially when the image is incidental.

These concerns motivate a more targeted question: when asked to judge a lexical property of a word, can a vision–language model (VLM) distinguish useful visual evidence from incidental image context? Concreteness and imagery ratings are well suited to this question because they reflect human judgments of general lexical properties: how much a word denotes something concrete, and how easily it evokes a mental image. Although these properties are perceptual in nature, the requested rating concerns the word itself rather than any particular picture. They therefore provide a controlled setting for testing whether VLMs preserve a lexical judgment when visual context is present. We use these ratings to test whether state-of-the-art VLMs preserve the requested word-level judgment when an image is present, including when the image is incidental to that judgment, or instead appear to treat retrieved images as direct evidence for a property of the word itself. We construct word–image pairs by varying the visual input while holding the target word and rating task fixed, comparing no-image input, uninformative placeholders, and *real images*, defined here as retrieved natural-image inputs. In summary, our contributions are as follows:

- We conduct a comprehensive evaluation of instruction-tuned text-only and vision–language models on human judgments of concreteness and

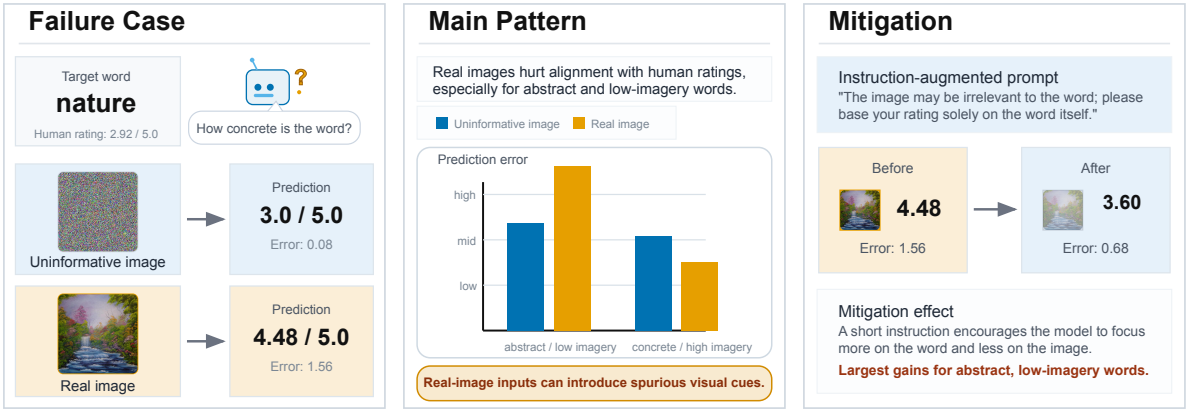



Figure 1: Real-image contexts can hurt alignment with human concreteness and imagery judgments, especially for abstract and low-imagery words. This pattern is associated with greater sensitivity to spurious visual cues, and instruction-augmented prompting reduces some errors by encouraging the model to base its rating on the word itself.

imagery across multiple visual contexts.

- We show that real-image contexts can hurt alignment with human ratings, especially for abstract and low-imagery words, and validate this contrast with a matched human study showing that human raters do not show comparable degradation under the same retrieved-image condition.
- We demonstrate a simple inference-time prompting intervention that reduces some of the degradation associated with real-image contexts.

## 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, often discussed as imageability, 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). The two properties are strongly correlated and often analyzed together, though they are not identical; the fine-grained distinction is not central to our diagnostic use of them. 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 principled diagnostic signals for studying perceptually grounded aspects of meaning. Motivated by prior work, we use these ratings to probe how modern multimodal models handle lexical judgments under visual context.

**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 models like LLaVA (Liu et al., 2023). A broad line of work shows that perceptual input can improve semantic representations, especially for concrete concepts, and that grounded semantic spaces often align with the concrete–abstract continuum (Hill and Korhonen, 2014; Zhuang et al., 2024b,a). However, recent work complicates this picture: text-only representations can outperform multimodal ones in capturing experiential semantic information and aligning with human fMRI responses (Bavaresco and Fernández, 2025), and ungrounded LLMs align better with human representations for non-sensorimotor than sensorimotor features of concepts (Xu et al., 2025).

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. We ask how different visual contexts affect VLM judgments of concreteness and imagery, and when they help or harm alignment with human ratings.

## 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 (Qwen Team, 2024). For VLMs, we evaluate LLaVA 1.5 (Liu et al., 2024), InternVL3 (Zhu et al., 2025), Pixtral (Agrawal et al., 2024), Qwen2.5-VL (Bai et al., 2025), and Gemma 3 (Gemma Team, 2025). All models are instruction-tuned. For model families where we evaluate both text-only and vision–language variants, 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 makes these comparisons more direct when contextualizing 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. For concreteness, we use a CP2004B-aligned subset of MT40k; for imagery, we use CP2004B entries with valid imagery ratings. After restricting both resources to single-word entries, the evaluation contains 2,082 MT40k words and 2,108 CP2004B 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. We assess retrieval quality through image–word relevance annotations on a CP2004B subset in Appendix C. In addition to these two real-image sources, we evaluate three controls: *None* (no image), *White* (a white image), and *Noise* (a random-noise image). 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. The intended judgment is lexical and metalinguistic: the model is asked to rate a general property of the word, not the concreteness or imagery of the accompanying image. Thus, the real-image contexts should be read as an incidental-context robustness test rather than as evidence that should determine the rating. To reduce prompt-template variation, we use five prompt variants per task and average the parsed predictions across them. This aggregation reduces prompt-template variance, but may hide prompt-specific behavior; we therefore report prompt-level variation in the representation analyses. VLMs receive the same textual prompts together with one of the five visual contexts (*None*, *White*, *Noise*, *ImageNet*, *Wikimedia*). We recognize that real-image contexts can nevertheless change the pragmatic frame of the prompt by making a visual instance salient. This ambiguity across word–image pairs is part of the phenomenon under study: a VLM may treat visual context as task-relevant even when the requested judgment is lexical. The *None*, *White*, and *Noise* controls help separate image presence and low-level visual input from real-image content; the matched human validation study in Appendix B provides the corresponding human comparison. The full prompt templates are listed in Appendix A.3.

### 3.4 Evaluation Methodology

**Model predictions.** For each model and each word–image pair, 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, low-imagery, and high-imagery subsets. Because RMSE penalizes deviations on the original rating scale, it measures whether predictions remain calibrated to human judgments under different visual contexts. 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 hidden states. This lets us test how strongly these properties are recoverable from hidden representations across network depth

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| **All** | | | | | |
| LLaMA 2 | 1.152 | – | – | – | – |
| Mistral | 0.803 | – | – | – | – |
| Qwen2.5 | 0.745 | – | – | – | – |
| Gemma 3 | **0.745** | $^{***}$0.933 | $^{***}$1.187 | $^{***}$0.787 | $^{***}$0.790 |
| 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 | 0.984 | $^{***}$1.085 | $^{***}$1.022 | $^{***}$0.943 | $^{***}$**0.931** |
| **Abstract** | | | | | |
| LLaMA 2 | 1.506 | – | – | – | – |
| Mistral | 0.939 | – | – | – | – |
| Qwen2.5 | 0.715 | – | – | – | – |
| Gemma 3 | 0.730 | $^{***}$**0.510** | $^{***}$0.607 | $^{***}$0.951 | $^{***}$0.932 |
| 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 | **0.765** | $^{***}$0.803 | 0.782 | $^{***}$1.069 | $^{***}$1.014 |
| **Concrete** | | | | | |
| LLaMA 2 | 0.728 | – | – | – | – |
| Mistral | 0.667 | – | – | – | – |
| Qwen2.5 | 0.769 | – | – | – | – |
| Gemma 3 | 0.758 | $^{***}$1.178 | $^{***}$1.513 | $^{***}$**0.616** | $^{***}$0.647 |
| 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 | 1.136 | $^{***}$1.277 | $^{***}$1.188 | $^{***}$**0.821** | $^{***}$0.855 |

MT40k (Concreteness)

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| **All** | | | | | |
| LLaMA 2 | 1.568 | – | – | – | – |
| Mistral | 1.392 | – | – | – | – |
| Qwen2.5 | 1.275 | – | – | – | – |
| Gemma 3 | **1.396** | $^{*}$1.420 | $^{***}$1.925 | $^{***}$2.046 | $^{***}$1.973 |
| 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 | **1.394** | $^{***}$1.440 | $^{***}$1.584 | $^{***}$1.782 | $^{***}$1.706 |
| **Low imagery** | | | | | |
| LLaMA 2 | 1.278 | – | – | – | – |
| Mistral | 1.819 | – | – | – | – |
| Qwen2.5 | 1.389 | – | – | – | – |
| Gemma 3 | 1.662 | $^{***}$1.611 | $^{***}$**0.804** | $^{***}$2.673 | $^{***}$2.567 |
| InternVL3 | 1.590 | $^{***}$1.372 | $^{***}$**0.739** | $^{***}$1.926 | $^{***}$1.842 |
| LLaVA 1.5 | **2.031** | $^{***}$2.627 | $^{***}$3.040 | $^{***}$3.066 | $^{***}$3.063 |
| Pixtral | 2.680 | $^{***}$**1.248** | $^{***}$1.867 | 2.666 | 2.709 |
| Qwen2.5-VL | **1.641** | 1.649 | $^{***}$1.796 | $^{***}$2.339 | $^{***}$2.239 |
| **High imagery** | | | | | |
| LLaMA 2 | 1.800 | – | – | – | – |
| Mistral | 0.809 | – | – | – | – |
| Qwen2.5 | 1.158 | – | – | – | – |
| Gemma 3 | **1.087** | $^{***}$1.213 | $^{***}$2.567 | $^{**}$1.181 | $^{*}$1.160 |
| InternVL3 | 0.918 | $^{***}$1.088 | $^{***}$2.838 | $^{***}$**0.815** | $^{***}$0.836 |
| LLaVA 1.5 | 1.163 | $^{***}$**0.903** | $^{***}$1.045 | $^{**}$1.061 | $^{***}$1.057 |
| Pixtral | **1.034** | $^{***}$1.534 | $^{***}$1.243 | $^{***}$1.173 | $^{***}$1.182 |
| Qwen2.5-VL | 1.113 | $^{***}$1.209 | $^{***}$1.354 | $^{**}$1.006 | $^{***}$**0.963** |

CP2004B (Imagery)

Table 1: RMSE of model predictions against human concreteness (left) and imagery (right) ratings. *None* denotes no image; *White*/*Noise* are uninformative contexts; *ImageNet*/*Wikimedia* are real-image contexts. Green/purple cells indicate better/worse within-row ranks; **boldface**/underlining mark best/worst contexts. Superscripts mark significance versus *None* for the same model ($^{*}$ $q < 0.05$, $^{**}$ $q < 0.01$, $^{***}$ $q < 0.001$; paired sign-flip tests with Benjamini–Hochberg correction).

under different visual contexts. For each layer, text-token representations are obtained by pooling over attended non-image tokens. For VLMs, we also probe image-token representations, pooled over model-specific image placeholder tokens, and combined representations, computed by averaging the pooled text-token and image-token representations. As a sanity check, we also report permuted-label probing baselines in Appendix E, 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 contexts, such as *None* versus real-image contexts. Given paired representation matrices $\mathbf{X}$ and $\mathbf{Z}$, where rows correspond to words and columns to hidden dimensions under two visual contexts, CCA finds projection vectors $\mathbf{a}$ and $\mathbf{b}$ that maximize

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

Here $\rho$ 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 contexts, whereas lower correlations indicate larger context-associated 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). 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 contexts. The full attribution formulation and implementation details are provided in Appendix A.8.

## 4 Results and Discussion

Our primary finding is that VLMs are vulnerable to visual cues that can be spurious for lexical judgments of concreteness and imagery. With real-image contexts, these models often show calibration loss and stronger sensitivity to visual content.

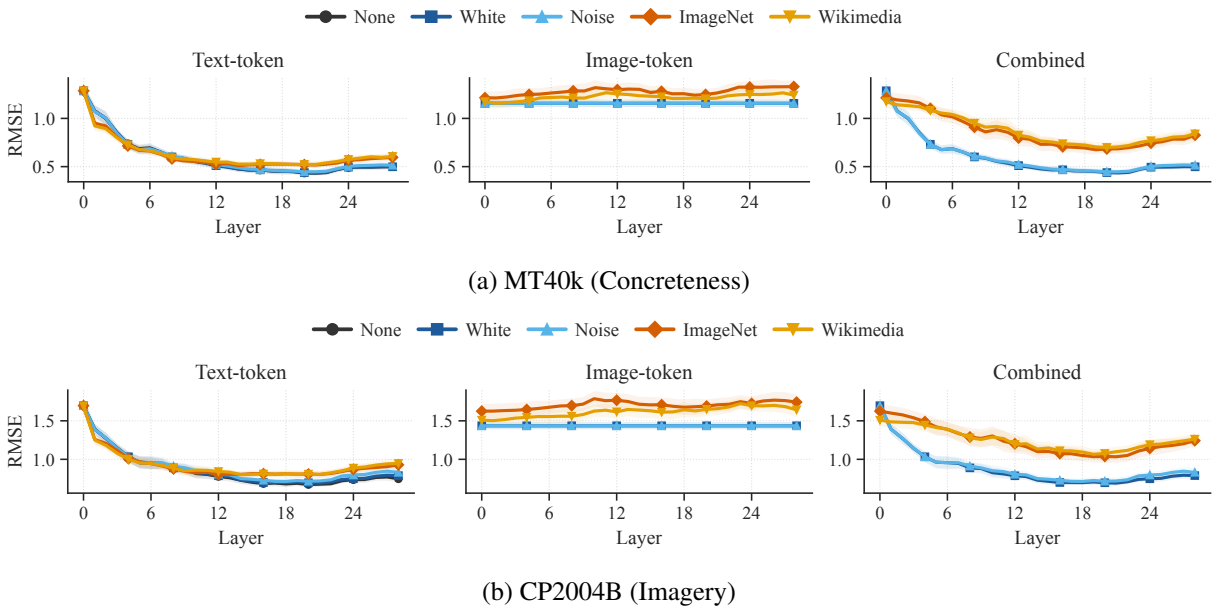



Figure 2: Layer-wise ridge probing for Qwen2.5-VL on (a) concreteness and (b) imagery. Panels group representation types; curves show visual contexts. Error bands show standard deviations across cv folds and prompt templates.

This is most evident on abstract and low-imagery words, where text-only models or VLMs given uninformative contexts often remain better calibrated. The following sections present detailed analyses supporting this conclusion.

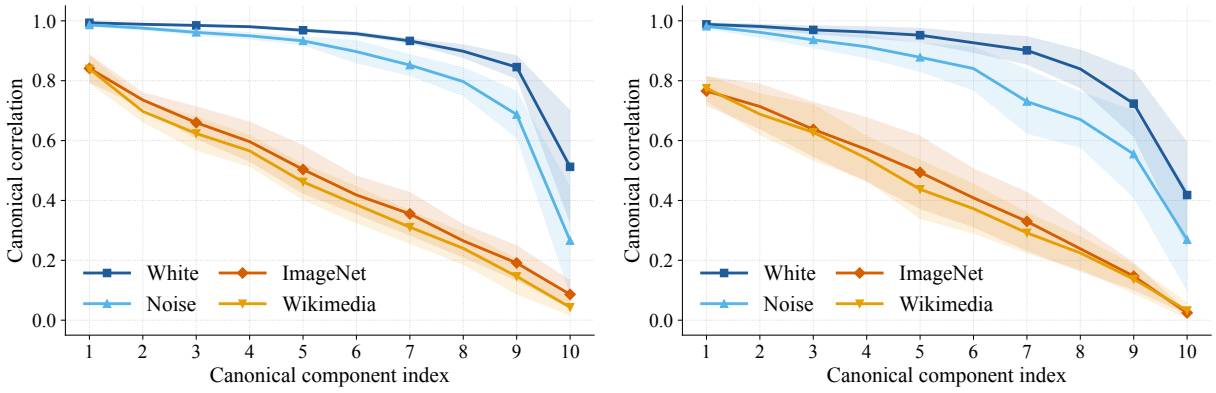
### 4.1 Model Predictions

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. Among VLMs, performance varies substantially based on the type of visual input provided. When real-image contexts are provided (from either *ImageNet* or *Wikimedia*), models often improve for concrete and high-imagery words but frequently worsen for abstract and low-imagery words. For abstract and low-imagery subsets, a *None* or *White* context yields lower RMSE than both real-image contexts for every VLM. Overall effects are therefore heterogeneous across models and subsets rather than uniformly positive or uniformly negative. Mean signed error analyses further show that real-image contexts often have the largest signed errors within a model/subset, reflecting an upward shift relative to other visual contexts (Appendix D.1). Spearman rank correlation results broadly support the RMSE pattern, while also showing that calibration and relative word ordering can diverge (Appendix D.2).

The improved performance on concrete and high-imagery words suggests that real-image contexts can be useful when the lexical target has clearer visual content, consistent with prior work on visual grounding for concrete concepts (Zhuang et al., 2024b,a). However, the substantial performance drop on abstract and low-imagery words supports the hypothesis that VLMs can be distracted by incidental visual cues in real-image contexts when image-word alignment is weak.

### 4.2 Human Validation

To test whether this pattern reflects VLM-specific sensitivity rather than a general pragmatic effect of making an image salient, Appendix B reports a matched human validation study on the CP2004B subset. Human raters also shift upward under retrieved images, indicating that images can influence lexical ratings in both humans and VLMs. However, humans do not show the same broad loss of calibration as VLMs, indicating that the VLM degradation cannot be reduced to the pragmatic effect of making an image salient during the judgment. Because retrieval quality is a potential confound, Appendix C further reports image–word relevance annotations for the same subset. After controlling for this confound, relevance-conditioned analyses still show that VLMs are significantly more sensitive than human raters to real-image con-




(a) MT40k (Concreteness)

(b) CP2004B (Imagery)

Figure 3: Canonical correlations between Qwen2.5-VL final-layer representations under *None* and each visual context for (a) concreteness and (b) imagery. Correlations use the top 10 PCA components; error bands show standard deviations across prompt templates.

text, as reflected in larger rating shifts.

### 4.3 Probing

Figure 2 shows layer-wise probing performance for concreteness and imagery for Qwen2.5 and Qwen2.5-VL under different visual contexts. For brevity, we present results using ridge regressors; results for other models, regressors, and metrics exhibit similar trends (see Appendix E). 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 under *None* or when the input image is uninformative (*White* or *Noise*), with probing curves closely matching those of text-only models.

However, using real-image contexts 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 real-image contexts also show reduced performance compared to text-only and uninformative contexts, indicating that these linguistic properties are less recoverable from text-token representations in real-image settings. When combining text and image hidden representations, performance remains in between that of text and image tokens alone, which is consistent with real-image contexts 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 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 contexts. Taken together, these results indicate that, in this setting, real-image contexts are associated with weaker recoverability of concreteness and imagery, especially when visual evidence is weakly aligned with the lexical target.

### 4.4 Canonical Correlation Analysis

Figure 3 presents the canonical correlations between the final-layer hidden representations obtained with different visual contexts for Qwen2.5-VL. The *None* visual context is compared against *White*, *Noise*, and real-image contexts. Hidden representations from uninformative contexts (*White* and *Noise*) exhibit high canonical correlations with *None* 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 from real-image contexts show markedly lower canonical correlations with *None*, with the last few components exhibiting correlations close to zero. This divergence indicates that real-image

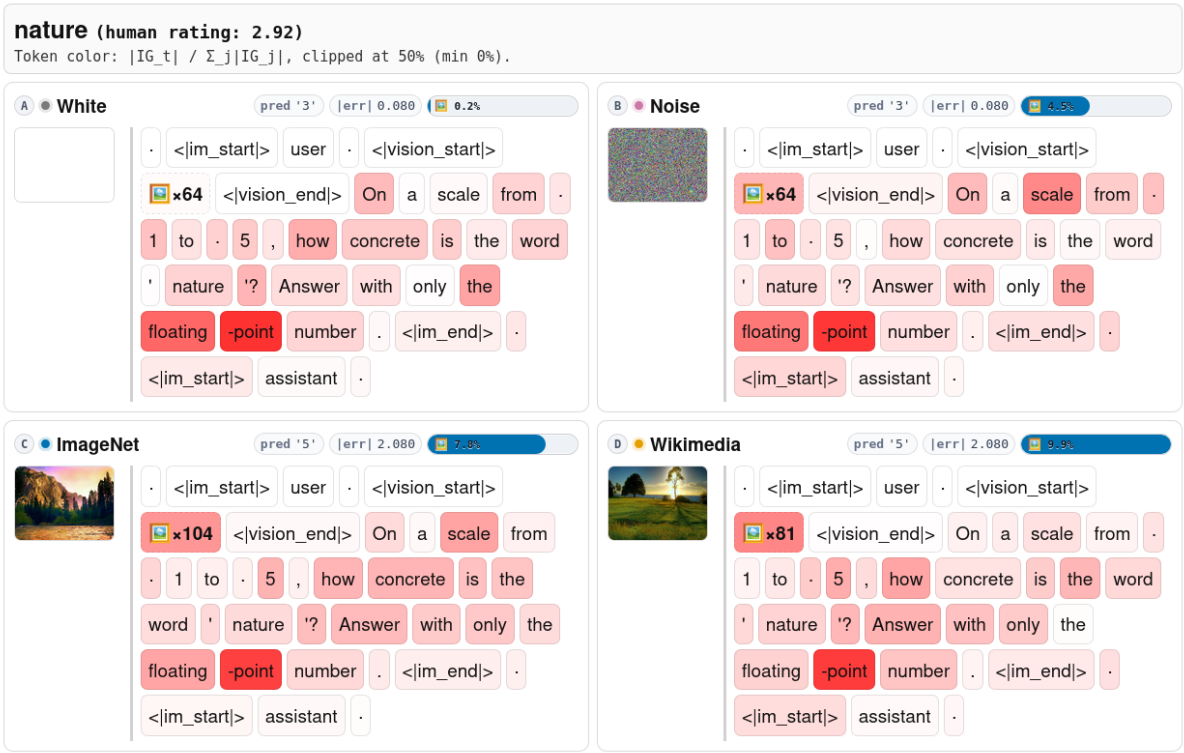



Figure 4: Integrated Gradients for Qwen2.5-VL on *"nature"* (human rating: 2.92) under four visual contexts. Panels report the predicted next token, absolute error, and image-token attribution share; token color shows normalized attribution magnitude.

contexts are associated with representational shifts relative to *None*, consistent with stronger sensitivity to visual content. These findings align with the probing results and suggest that real-image contexts can reduce agreement with human concreteness and imagery ratings for this task.

### 4.5 Interpretability: A Case Study

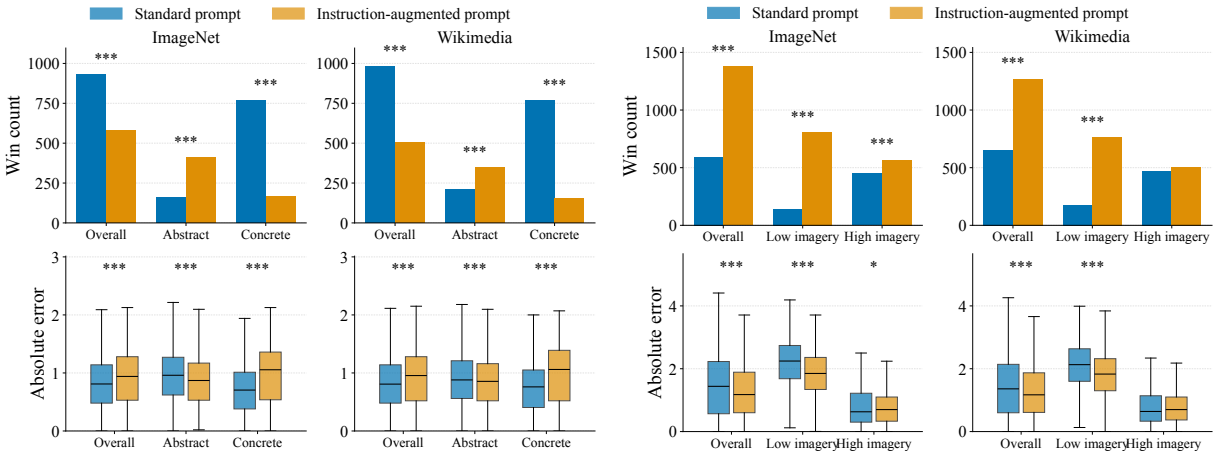
Figure 4 presents a case study of Qwen2.5-VL on the target word *nature* (human concreteness rating: 2.92) using LayerIG. The example shows how attribution shifts from text tokens toward image tokens as the input changes from uninformative to real-image contexts, and how this shift tracks changes in prediction error.

With *White* and *Noise* contexts, attributions are concentrated on text tokens, and the model predicts 3, close to the human rating (absolute error 0.08). This indicates that when visual input is uninformative, the model relies primarily on linguistic cues and remains relatively well calibrated.

However, in real-image contexts (*ImageNet* and *Wikimedia*), the attribution landscape shifts. Image-token attribution rises from 0.2% for *White* and 4.5% for *Noise* to 7.8% for *ImageNet* and 9.9% for *Wikimedia*, while the predicted token shifts upward to 5 in both real-image contexts (absolute error 2.08). Together, these changes suggest greater image reliance and weaker calibration. The resulting overestimation is consistent with our broader finding that real-image contexts are associated with upward-biased concreteness judgments when visual evidence is incidental to the lexical target.

## 5 Mitigation Strategies

Our preceding analyses showed that real-image contexts are associated with worse alignment with human ratings, shifted hidden representations, and higher attribution to image tokens in a case study, especially when the visual content is incidental to the target word. One plausible interpretation is that the model exhibits a form of shortcut learning at inference time: once a real-image context 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 partly depends on how the model interprets the prompt, then making the possible irrelevance of the image explicit may help it discount misleading visual evidence.




(a) MT40k (Concreteness)

(b) CP2004B (Imagery)

Figure 5: Instruction-augmented prompting for Qwen2.5-VL in real-image contexts on (a) concreteness and (b) imagery. Top rows show non-tied win counts, excluding tied absolute errors; bottom rows show absolute-error distributions over all valid paired predictions, overall and by rating subset. Superscripts mark significance versus standard prompting after Benjamini–Hochberg correction: exact sign tests for non-tied win counts and paired sign-flip tests for absolute-error differences ($^{*}$ $q < 0.05$, $^{**}$ $q < 0.01$, $^{***}$ $q < 0.001$).

We therefore test a simple prompting intervention with Qwen2.5-VL as the representative model. For each of the five original rating prompt variants, we prepend the same sentence:

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

Following the instructed-prompting approach of Shi et al. (2023), this intervention 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 under both real-image contexts using non-tied win counts and absolute-error distributions. The effect is selective rather than uniform. On MT40k, the instruction-augmented prompt wins frequently on abstract words, but gains are small in magnitude and the concrete subset still favors the standard prompt. On CP2004B, the instruction-augmented prompt is preferred overall and especially for low-imagery words, with visibly lower error distributions in those subsets; by contrast, the high-imagery subset is close to parity and shows little magnitude difference between prompts. Taken together, these results indicate that even this single instruction can reduce some errors in real-image contexts, primarily for subsets most vulnerable to real-image-associated overestimation. Appendix F extends this analysis to other VLMs and tests five Qwen2.5-VL mitigation instruction variants. These extensions show that the mitigation pattern is not unique to Qwen2.5-VL: several other VLMs also benefit, with the most consistent gains on CP2004B low-imagery words, while Qwen2.5-VL wording variants preserve the same selective tradeoff between improved abstract or low-imagery subsets and weaker concrete-word results.

Because this intervention operates entirely at inference time, it offers a lightweight guardrail for noisy, weakly aligned, or incidental visual contexts, though stronger mitigation will likely benefit from methods that adaptively calibrate visual evidence.

## 6 Conclusion

We show that real-image contexts can undermine VLM alignment with human concreteness and imagery judgments, especially for abstract and low-imagery words. Crucially, a matched human study shows no comparable degradation, indicating that the VLM pattern is not explained solely by the pragmatic effect of making an image salient. Our comprehensive analysis relates this degradation to weaker recoverability of the targeted lexical properties in probing and context-associated representational shifts. A simple instruction to base the rating on the word itself reduces some errors in these vulnerable subsets. These findings suggest that current instruction-tuned VLMs would benefit from better calibration of when visual context is relevant to the requested linguistic judgment.

## Limitations

Our study has two main limitations. First, evaluation is limited to English lexical norms (MT40k and CP2004B) and single-word items, so findings may not transfer directly to larger lexical resources, other languages, multi-word expressions, or sentence-level grounding. Second, although Appendix F.2 tests five mitigation instruction variants for Qwen2.5-VL, our mitigation study remains a lightweight prompt-level analysis rather than a full search over mitigation designs. The selective gains therefore should not be read as the limit of possible mitigation; stronger approaches may require model- or task-specific calibration strategies.

## 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*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Théophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *arXiv preprint arXiv: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*.

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 arXiv:2502.13923*.

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

Anna Bavaresco and Raquel Fernández. 2025. Experiential semantic information and brain alignment: Are multimodal models better than language models? In *Proceedings of the 29th Conference on Computational Natural Language Learning*.

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

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

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*.

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*.

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*.

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*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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*.

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)*.

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)*.

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*.

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*.

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*, Proceedings of Machine Learning Research.

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*.

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*.

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)*.

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

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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*, Proceedings of Machine Learning Research.

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*, Proceedings of Machine Learning Research.

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*.

Richard D. Sperber, Charley McCauley, Ronnie D. Ragain, and Carolyne M. Weil. 1979. Semantic priming effects on picture and word processing. *Memory & Cognition*, 7:339–345.

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*, Proceedings of Machine Learning Research.

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*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan 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 arXiv: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*.

Qihui Xu, Yingying Peng, Samuel A. Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*, 9:1871–1886.

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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024a. Lexicon-level contrastive visual-grounding improves language modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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)*.

# 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 Source Data, Filtering, and Licenses

We construct task-specific single-word evaluation sets from MT40k and CP2004B. For MT40k, we start from a CP2004B-aligned file and drop missing ratings, bigrams, hyphenated forms, and duplicate lowercased surface forms. For CP2004B, we retain entries with valid positive imagery ratings and apply the same normalization, hyphen filtering, and deduplication. Table 3 summarizes the resulting rating distributions.

| Dataset | Words | Mean | SD | Range | Unique | Med. gap |
|---|---|---|---|---|---|---|
| MT40k | 2,082 | 3.29 | 1.16 | 1.07–5.00 | 326 | 0.01 |
| CP2004B | 2,108 | 4.21 | 1.44 | 1.16–6.90 | 295 | 0.02 |

Table 3: Rating statistics for the processed single-word evaluation files. *Unique*: number of distinct rating values; *Med. gap*: median positive gap between adjacent values.

MT40k is distributed under CC BY-NC-ND 3.0,[1] while CP2004B remains under Psychonomic archive terms that retain author rights and permit non-commercial research or educational use with citation.[2] We do not redistribute original or processed item-level rating tables; instead, we report summary statistics and aggregate results and provide preprocessing code for locally obtained sources. Retrieved images remain subject to ImageNet and Wikimedia Commons source terms.[3][4]

[1] https://rdrr.io/github/JackEdTaylor/LexOPS/f/inst/extdata/data-licenses/BRYSBAERT-ET-AL-2014.md

[2] https://rdrr.io/github/JackEdTaylor/LexOPS/f/inst/extdata/data-licenses/CLARK-AND-PAIVIO-LICENSE.md

[3] https://www.image-net.org/download.php

[4] https://commons.wikimedia.org/wiki/Commons:Licensing

We release anonymized human-study ratings and newly produced relevance annotations under CC BY 4.0, excluding Prolific/session identifiers and raw third-party images.[5]

## 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.*

[5] https://creativecommons.org/licenses/by/4.0/

### A.4 Visual Input Contexts

We evaluate the same five visual contexts used in the main paper: *None*, *White*, *Noise*, *ImageNet*, and *Wikimedia*. In the pipeline, the *Noise* context is implemented as a random-noise placeholder image. For real-image contexts, each word is associated with a directory of up to five retrieved images, and image assignment is shuffled before prompt-template pairing. This produces multiple word–image pairs for the same lexical target while keeping the rating task fixed.

### 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. For each layer and prompt template, text-token representations are computed by mean pooling hidden states over all attended non-image tokens after chat templating, and image-token representations are computed by mean pooling hidden states over model-specific image placeholder tokens. The combined representation is the average of the pooled text-token and image-token vectors; when no image tokens are present, only text-token representations are used. Curves report the mean across cross-validation folds and prompt templates, with standard deviations shown as error bands. 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 $1 \times 10^{-5}$). 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 context and align words across the two contexts being compared (e.g., *None* vs. *White*, *Noise*, or a real-image context). 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:

$$\rho = \max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^\top \Sigma_{XZ} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{XX} \mathbf{a}\, \mathbf{b}^\top \Sigma_{ZZ} \mathbf{b}}}, \qquad (2)$$

Here $\rho$ denotes the resulting canonical correlation, and $\Sigma_{XX}$, $\Sigma_{ZZ}$, and $\Sigma_{XZ}$ are the empirical covariance matrices and cross-covariance matrix. 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

$$\mathrm{IG}_{t,k}(\mathbf{x}) = (e_{t,k} - e'_{t,k}) \int_0^1 \frac{\partial F\big(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\big)}{\partial e_{t,k}} \, d\alpha. \tag{3}$$

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

$$\mathrm{IG}_t = \sum_{k=1}^{d} \mathrm{IG}_{t,k}. \tag{4}$$

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

$$A_{\text{modality}} = \sum_{t \in \mathcal{T}_{\text{modality}}} |\mathrm{IG}_t|, \tag{5}$$

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 Human Validation Experiments

### B.1 Experiment Setup

To check whether the real-image rating shift observed in VLMs also appears in human ratings, we ran a human validation study on a 500-word subset of CP2004B. The subset was sampled evenly across ten CP2004B imagery deciles, so that both low- and high-imagery words were covered. Participants were recruited through Prolific using screening criteria for English as a first language and residence in the United States, United Kingdom, Canada, Australia, New Zealand, or Ireland. They rated target words on the CP2004B 1–7 imagery scale.

The study compared two visual contexts from the model analysis: the uninformative *White* context and the retrieved *ImageNet* context. Fifty participants each completed 100 trials, with an equal number of *White* and *ImageNet* trials and no repeated word within a participant. Across the full study, each word–image pair received five ratings. We averaged human ratings at the word–image level and compared them with both the published CP2004B ratings and the corresponding VLM predictions on the same 500-word subset.

The study was configured for an estimated duration of 8–10 minutes and a maximum allowed time of 15 minutes; the observed mean completion time was 6.68 minutes (median 5.98 minutes) for the 50 completed sessions. Prolific reported an average reward rate of £12.10 per hour, above common minimum-wage thresholds. The task collected ratings and Prolific URL identifiers for submission tracking; no free-text responses or other personal data were requested. The data collection protocol was approved by an institutional research ethics review board.

### B.2 Participant Instructions

Before the instructions, participants viewed a consent page stating the expected study duration and that Prolific submission information would be recorded; they proceeded only after indicating consent. They were then warned that the task could include potentially disturbing images and unpleasant words, and were informed that they would rate about 100 English words. The instruction page then presented the question "How easily and quickly does this word arouse a mental image?" It defined imagery as follows: "Words differ in how easily they arouse mental images of things or events. A mental image can be a mental picture, sound, or other sensory experience. Give a high rating to words that arouse a mental image quickly and easily. Give a low rating to words that arouse a mental image only with difficulty, after a delay, or not at all. Use the scale to indicate the ease or difficulty with which each word arouses a mental image for you." The same page explained that the visual display could be a blank image or an image from an image dataset. The response scale was shown as 1 = very difficult to image, 2 = difficult, 3 = somewhat difficult, 4 = moderate, 5 = somewhat easy, 6 = easy, and 7 = very easy to image.

### B.3 Results and Discussion

Table 6 compares the averaged human ratings with VLM predictions on the same 500-word subset.

The central comparison is the change from the uninformative *White* context to the retrieved *ImageNet* context. For RMSE, humans show only a small overall increase, whereas every VLM shows a larger and significant increase. The contrast is clearest for low-imagery words: human ratings move modestly away from the published CP2004B ratings, but several VLMs show much larger losses in calibration. The $\Delta - \Delta_H$ columns confirm this contrast statistically: for the full subset and the low-imagery subset, every VLM's RMSE increase is significantly larger than the corresponding human increase. For high-imagery words, humans and several VLMs instead improve under *ImageNet*, consistent with real-image contexts being more useful when the word has clearer visual content.

Mean signed error shows the direction of the shift. Both humans and VLMs assign higher imagery ratings under *ImageNet*, consistent with prior evidence that semantic context can influence both picture and word processing (Sperber et al., 1979). However, the upward shift is generally larger for VLMs, especially for low-imagery words. The model–human difference in signed-error shift is significant for most VLMs across these splits, indicating that the retrieved images produce a stronger upward bias in model ratings than in human ratings. Spearman rank correlation adds a complementary view: rank agreement can improve even when RMSE and signed error worsen, showing that visual context may preserve or sharpen item ordering while still shifting predictions away from the human rating scale. The human comparison shows that the VLM degradation cannot be explained by the pragmatic effect of making an image salient during the judgment alone; instead, VLMs are more sensitive than human raters to incidental visual content when making the lexical judgment.

## C Image-Word Relevance Annotation

To assess retrieval quality as a potential confound in the CP2004B results, we manually annotated image–word relevance for the same 500-word subset used in the human validation study. For each word, we reproducibly sampled one of the five *ImageNet* images used in the human study. Three annotators independently rated each word–image pair on a three-point scale: 0 for unrelated or misleading, 1 for weak, indirect, or ambiguous, and 2 for a clear visual match. Each of the 500 pairs received three ratings. Inter-annotator agreement was moderate (Fleiss' $\kappa = 0.40$); exact three-way agreement was 46.2%, and mean pairwise agreement was 62.5%. All three annotators were authors of this paper.

Table 4 summarizes the majority relevance labels. Retrieved images were clear matches for most high-imagery words (72.3%), but for only 20.9% of low-imagery words. Conversely, 26.6% of low-imagery words were judged unrelated or misleading, compared with 3.9% of high-imagery words. This confirms that image–word match quality is strongly tied to the target lexical property: low-imagery words are exactly the cases where retrieved visual evidence is least reliable.

| Majority label | All | Low imagery | High imagery |
|---|---|---|---|
| Unrelated | 75 (15.0%) | 65 (26.6%) | 10 (3.9%) |
| Weak/ambig. | 189 (37.8%) | 128 (52.5%) | 61 (23.8%) |
| Clear match | 236 (47.2%) | 51 (20.9%) | 185 (72.3%) |

Table 4: Image–word relevance annotations for one sampled *ImageNet* image per word in the 500-word CP2004B subset. Labels are majority votes over three annotators.

To control for this retrieval-quality confound, we also conditioned the *ImageNet*-minus-*White* shift on the majority relevance labels (Table 5). Human ratings show only a small shift for unrelated images (+0.05), whereas VLMs generally shift upward even when the image is judged unrelated or misleading. Many VLM shifts are significantly larger than the corresponding human shifts, including for unrelated images. Because these comparisons are made within each relevance category, the larger VLM shifts cannot be attributed only to poorer image matches for low-imagery words. Even clear matches are often associated with larger model than human rating shifts, showing that relevance helps explain part of the variation but does not eliminate the VLM-specific sensitivity to visual context. Thus, retrieval quality is a meaningful source of variation, but the relevance-conditioned analysis supports the main conclusion: VLMs are more sensitive than human raters to incidental visual content in this lexical judgment task.

| Source | Unrel. | Weak | Clear |
|---|---|---|---|
| Human | **+0.05** | **+0.38** | **+0.51** |
| Gemma 3 | $^{***}$+0.83 | $^{***}$+1.22 | $^{***}$+1.19 |
| InternVL3 | +0.30 | $^{*}$+0.61 | $^{***}$+0.90 |
| LLaVA 1.5 | $^{*}$+0.43 | +0.53 | +0.52 |
| Pixtral | $^{***}$+1.65 | $^{***}$+1.65 | $^{***}$+1.91 |
| Qwen2.5-VL | $^{*}$+0.44 | $^{***}$+0.87 | $^{***}$+1.33 |

Table 5: Mean *ImageNet*-minus-*White* rating shifts on the 500-word CP2004B subset by majority image–word relevance label. Positive values indicate higher ratings under *ImageNet*. Star superscripts mark paired significance versus the human shift after Benjamini–Hochberg correction within each relevance label ($^{*}$ $q < 0.05$, $^{**}$ $q < 0.01$, $^{***}$ $q < 0.001$).

# D Additional Model Prediction Results

## D.1 Signed Prediction Errors

Table 7 reports the direction of prediction error relative to human ratings. Positive values indicate overestimation, while negative values indicate underestimation. The relevant pattern is comparative: real-image contexts often have the most positive signed errors within a model/subset row, indicating an upward shift relative to the other visual contexts. For some rows, this shift produces positive overestimation; for others, predictions remain below the human ratings in every context, but the real-image contexts are closest to zero. This pattern is strongest for imagery and is also visible for abstract words, supporting the main RMSE results: real-image contexts can push ratings upward when the requested judgment should remain lexical.

## D.2 Spearman Rank Correlations

The main results use RMSE because our primary question is whether visual context moves predictions away from the human rating scale. Table 8 adds the corresponding Spearman rank correlations for the same models, datasets, and visual contexts. This metric asks a different question: whether the relative ordering of words is preserved, even if the predicted scale shifts. We therefore treat Spearman as complementary, especially because the rating scales are dense; small swaps among similarly rated words can affect rank-based scores, and Table 3 documents this close spacing through the median gap between adjacent distinct ratings.

The Spearman results mostly support the RMSE analysis, but with one useful distinction. *None* and *White* remain strong overall, consistent with the main finding that removing or minimizing visual content generally preserves lexical judgments. At the same time, *Noise* is more damaging under Spearman than under RMSE for several models, suggesting that predictions can stay near the correct scale while losing some item-level ordering. This separation between calibration and rank order also appears in the human validation study (Appendix B), where rank agreement can improve even when RMSE and mean signed error show a larger real-image rating shift. Taken together, the Spearman results refine the RMSE picture: they show when item-level ordering and scale calibration diverge, while leaving intact the main conclusion that visual context is associated with substantial shifts in model predictions for lexical judgments.

# E Additional Probing Results

This section provides additional probing results corresponding to the analyses summarized in the main text. For each dataset (MT40k and CP2004B), we report layer-wise probing curves for all three regressors (ridge, SVM, and MLP) under RMSE and Spearman rank correlation, together with matched permuted-label baselines. Figures 6–8 show the MT40k results, and Figures 9–11 show the CP2004B results.

# F Additional Mitigation Analyses

## F.1 Additional Model Results

This section provides additional mitigation results for the other evaluated VLMs under the two real-image contexts. As in the main experiment, the intervention prepends the same sentence to each of the five original rating prompt variants. As in Figure 5, non-tied win counts exclude tied absolute errors, while the lower panels show absolute-error distributions over all valid paired predictions; superscripts use exact sign tests for non-tied win counts and paired sign-flip tests for absolute-error differences. Figure 12 reports MT40k results, and Figure 13 reports CP2004B results. Overall, the additional models support the main mitigation result: making image irrelevance explicit can reduce errors, most consistently on CP2004B low-imagery words. The effect is less uniform on MT40k, where gains depend more on the model and subset, so the intervention is best viewed as a lightweight, selective mitigation rather than a universal fix.

### F.2 Mitigation Prompt Wording Robustness

To test whether the Qwen2.5-VL mitigation results depend on instruction wording, we keep the model, visual contexts, and original rating prompt variants fixed while varying only the added instruction. Predictions are averaged across the five original rating prompt variants, as in the main experiments. Alongside the instruction used in the main experiment, we test four related variants:

1. *Ignore visual details that are unrelated to the target word, and rate the word based on its meaning.*
2. *Use the image only if it helps with the target word; otherwise, rely on the word itself.*
3. *Do not let irrelevant image content influence your rating of the word.*
4. *If the image and word seem mismatched, prioritize the word when assigning the rating.*

Table 9 summarizes the results. *Win margin* is the mitigation win count minus the standard-prompt win count, divided by the number of non-tied pairs and reported in percentage points. *MAE reduction* is the mean item-level decrease in absolute error, $|\hat{y}_{\text{standard}} - y| - |\hat{y}_{\text{mitigation}} - y|$, so positive values favor the mitigation prompt.

On MT40k, the variants reproduce the subset-level tradeoff seen in the main mitigation result. The average rows show consistent gains for abstract words in both real-image contexts, while concrete words move in the opposite direction. The overall MT40k averages are therefore negative, indicating that the wording variants improve the abstract subset but do not provide a uniform benefit across the dataset. CP2004B shows broader support for mitigation. The average rows are positive overall, with the clearest improvements on low-imagery words. High-imagery words show little average change, suggesting that the mitigation mainly helps where irrelevant visual evidence is most disruptive. Overall, these results suggest that the gains are robust across prompt wordings for the cases most vulnerable to irrelevant visual evidence, but the benefit remains selective rather than universal.

| Source | All | | | | Low imagery | | | | High imagery | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ |
| Human | **1.15** | **1.20** | **+0.04** | – | 1.24 | **1.43** | $^{**}$**+0.19** | – | 1.07 | 0.93 | $^{*}$-0.14 | – |
| Gemma 3 | 1.41 | 2.03 | $^{***}$+0.62 | $^{***}$+0.58 | 1.58 | 2.65 | $^{***}$+1.08 | $^{***}$+0.89 | 1.23 | 1.17 | -0.06 | **+0.08** |
| InternVL3 | 1.20 | 1.43 | $^{***}$+0.23 | $^{***}$**+0.19** | 1.33 | 1.89 | $^{***}$+0.55 | $^{***}$+0.36 | 1.06 | **0.78** | $^{***}$-0.28 | -0.14 |
| LLaVA 1.5 | 1.93 | 2.26 | $^{***}$+0.33 | $^{***}$+0.29 | 2.61 | 3.05 | $^{***}$+0.44 | $^{***}$**+0.25** | **0.87** | 1.05 | $^{***}$+0.17 | $^{***}$+0.32 |
| Pixtral | 1.36 | 2.02 | $^{***}$+0.65 | $^{***}$+0.61 | **1.19** | 2.63 | $^{***}$+1.44 | $^{***}$+1.25 | 1.51 | 1.16 | $^{***}$**-0.35** | $^{*}$-0.21 |
| Qwen2.5-VL | 1.42 | 1.76 | $^{***}$+0.34 | $^{***}$+0.30 | 1.62 | 2.32 | $^{***}$+0.70 | $^{***}$+0.51 | 1.19 | 0.97 | $^{**}$-0.23 | **-0.08** |

(a) RMSE

| Source | All | | | | Low imagery | | | | High imagery | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ |
| Human | +0.25 | **+0.64** | $^{***}$**+0.39** | – | **+0.73** | **+1.02** | $^{***}$**+0.29** | – | -0.21 | +0.28 | $^{***}$**+0.49** | – |
| Gemma 3 | +0.45 | +1.59 | $^{***}$+1.15 | $^{***}$+0.75 | +1.40 | +2.54 | $^{***}$+1.14 | $^{***}$+0.85 | -0.46 | +0.69 | $^{***}$+1.16 | $^{***}$+0.66 |
| InternVL3 | +0.19 | +0.89 | $^{***}$+0.70 | $^{***}$+0.31 | +1.18 | +1.72 | $^{***}$+0.54 | $^{**}$+0.25 | -0.76 | **+0.09** | $^{***}$+0.85 | $^{***}$+0.36 |
| LLaVA 1.5 | +1.26 | +1.76 | $^{***}$+0.51 | $^{*}$**+0.11** | +2.49 | +2.98 | $^{***}$+0.48 | $^{*}$**+0.19** | **+0.08** | +0.61 | $^{***}$+0.53 | **+0.04** |
| Pixtral | **-0.13** | +1.65 | $^{***}$+1.77 | $^{***}$+1.38 | +0.83 | +2.53 | $^{***}$+1.70 | $^{***}$+1.41 | -1.03 | +0.80 | $^{***}$+1.84 | $^{***}$+1.34 |
| Qwen2.5-VL | +0.30 | +1.33 | $^{***}$+1.02 | $^{***}$+0.63 | +1.48 | +2.19 | $^{***}$+0.71 | $^{***}$+0.42 | -0.82 | +0.50 | $^{***}$+1.32 | $^{***}$+0.83 |

(b) Mean signed error

| Source | All | | | | Low imagery | | | | High imagery | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ | *White* | *ImageNet* | $\Delta$ | $\Delta - \Delta_H$ |
| Human | **0.67** | **0.75** | $^{**}$+0.08 | – | 0.40 | 0.44 | +0.03 | – | **0.51** | **0.60** | +0.09 | – |
| Gemma 3 | 0.41 | 0.44 | +0.03 | -0.05 | **0.53** | **0.49** | -0.04 | -0.07 | 0.08 | 0.10 | +0.02 | -0.07 |
| InternVL3 | 0.59 | 0.64 | +0.05 | **-0.04** | 0.33 | 0.32 | -0.02 | **-0.05** | 0.48 | 0.48 | -0.01 | -0.10 |
| LLaVA 1.5 | -0.01 | 0.13 | +0.14 | +0.06 | -0.09 | 0.07 | +0.17 | $^{*}$+0.13 | 0.08 | 0.03 | -0.05 | -0.14 |
| Pixtral | 0.39 | 0.59 | $^{**}$+0.21 | +0.12 | 0.40 | 0.48 | +0.08 | **+0.05** | 0.12 | 0.25 | +0.14 | **+0.04** |
| Qwen2.5-VL | 0.18 | 0.58 | $^{**}$**+0.40** | $^{***}$+0.32 | 0.13 | 0.42 | $^{**}$**+0.29** | $^{*}$+0.26 | 0.10 | 0.26 | **+0.15** | +0.06 |

(c) Spearman rank correlation

Table 6: Human validation metrics on the CP2004B subset under *White* and retrieved *ImageNet* contexts, with $\Delta = \textit{ImageNet} - \textit{White}$. Positive $\Delta$ means worse RMSE, upward signed-error shift, or improved Spearman rank agreement. $\Delta - \Delta_H$ is the source-specific $\Delta$ minus the human $\Delta$ for the same split and metric. **Boldface**/underlining mark best/worst values per metric column. Star superscripts on $\Delta$ mark paired significance for *ImageNet* versus *White* within a source; star superscripts on $\Delta - \Delta_H$ mark paired significance versus the human $\Delta$, after Benjamini–Hochberg correction ($^{*}$ $q < 0.05$, $^{**}$ $q < 0.01$, $^{***}$ $q < 0.001$).

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| **All** | | | | | |
| Gemma 3 | **-0.02** | -0.42 | -0.54 | +0.21 | +0.17 |
| InternVL3 | -0.31 | -0.49 | -1.17 | **-0.01** | -0.05 |
| LLaVA 1.5 | +0.75 | **+0.71** | +0.71 | +0.71 | +0.71 |
| Pixtral | +0.59 | **-0.30** | +0.31 | +0.46 | +0.46 |
| Qwen2.5-VL | -0.28 | -0.33 | -0.31 | +0.10 | **+0.05** |
| **Abstract** | | | | | |
| Gemma 3 | +0.62 | **+0.30** | +0.41 | +0.80 | +0.78 |
| InternVL3 | +0.45 | +0.32 | **-0.04** | +0.68 | +0.65 |
| LLaVA 1.5 | +1.86 | **+1.82** | +1.82 | +1.82 | +1.82 |
| Pixtral | +1.50 | **+0.37** | +1.27 | +1.14 | +1.12 |
| Qwen2.5-VL | **+0.60** | +0.66 | +0.61 | +0.95 | +0.89 |
| **Concrete** | | | | | |
| Gemma 3 | -0.56 | -1.03 | -1.35 | **-0.29** | -0.35 |
| InternVL3 | -0.96 | -1.18 | -2.13 | **-0.59** | -0.64 |
| LLaVA 1.5 | **-0.19** | -0.23 | -0.23 | -0.23 | -0.23 |
| Pixtral | -0.18 | -0.87 | -0.50 | -0.12 | **-0.10** |
| Qwen2.5-VL | -1.01 | -1.16 | -1.08 | **-0.61** | -0.66 |

MT40k (Concreteness)

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| **All** | | | | | |
| Gemma 3 | +0.54 | **+0.48** | -1.11 | +1.59 | +1.49 |
| InternVL3 | +0.49 | **+0.20** | -1.38 | +0.91 | +0.81 |
| LLaVA 1.5 | **+0.65** | +1.26 | +1.72 | +1.77 | +1.76 |
| Pixtral | +1.55 | **-0.10** | +0.41 | +1.65 | +1.69 |
| Qwen2.5-VL | +0.39 | +0.33 | **+0.31** | +1.33 | +1.21 |
| **Low imagery** | | | | | |
| Gemma 3 | +1.51 | +1.42 | **+0.18** | +2.56 | +2.45 |
| InternVL3 | +1.48 | +1.20 | **-0.05** | +1.76 | +1.67 |
| LLaVA 1.5 | **+1.88** | +2.51 | +2.95 | +2.99 | +2.99 |
| Pixtral | +2.62 | **+0.86** | +1.67 | +2.56 | +2.61 |
| Qwen2.5-VL | +1.52 | **+1.51** | +1.55 | +2.20 | +2.11 |
| **High imagery** | | | | | |
| Gemma 3 | **-0.37** | -0.40 | -2.32 | +0.68 | +0.59 |
| InternVL3 | -0.45 | -0.74 | -2.64 | +0.10 | **+0.01** |
| LLaVA 1.5 | -0.51 | **+0.08** | +0.57 | +0.61 | +0.61 |
| Pixtral | **+0.55** | -1.00 | -0.77 | +0.79 | +0.82 |
| Qwen2.5-VL | -0.67 | -0.78 | -0.86 | +0.51 | **+0.37** |

CP2004B (Imagery)

Table 7: Mean signed prediction error, $\hat{y}_{\text{context}} - y_{\text{human}}$, for concreteness (left) and imagery (right). Positive/negative values indicate over-/underestimation. *White*/*Noise* are uninformative contexts; *ImageNet*/*Wikimedia* are real-image contexts. Green/purple cells mark smaller/larger within-row absolute errors; **boldface**/underlining mark best/worst contexts.

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| | | | **All** | | |
| LLaMA 2 | 0.426 | – | – | – | – |
| Mistral | 0.737 | – | – | – | – |
| Qwen2.5 | 0.829 | – | – | – | – |
| Gemma 3 | **0.797** | $^{***}$0.730 | $^{***}$0.430 | $^{*}$0.775 | $^{**}$0.767 |
| InternVL3 | **0.775** | $^{*}$0.751 | $^{***}$-0.111 | $^{*}$0.744 | $^{**}$0.736 |
| LLaVA 1.5 | 0.039 | **0.040** | $^{**}$-0.085 | 0.004 | 0.022 |
| Pixtral | 0.659 | 0.655 | $^{***}$0.410 | 0.655 | $^{*}$**0.695** |
| Qwen2.5-VL | 0.600 | $^{***}$0.497 | $^{*}$0.564 | 0.629 | $^{*}$**0.642** |
| | | | **Abstract** | | |
| LLaMA 2 | 0.146 | – | – | – | – |
| Mistral | 0.281 | – | – | – | – |
| Qwen2.5 | 0.358 | – | – | – | – |
| Gemma 3 | **0.413** | 0.346 | $^{**}$0.214 | 0.358 | 0.356 |
| InternVL3 | **0.274** | 0.264 | $^{**}$0.036 | 0.238 | 0.244 |
| LLaVA 1.5 | -0.039 | 0.000 | 0.025 | **0.036** | -0.004 |
| Pixtral | 0.229 | 0.224 | 0.206 | 0.234 | **0.287** |
| Qwen2.5-VL | 0.172 | 0.117 | $^{*}$0.070 | 0.154 | **0.199** |
| | | | **Concrete** | | |
| LLaMA 2 | 0.286 | – | – | – | – |
| Mistral | 0.640 | – | – | – | – |
| Qwen2.5 | 0.729 | – | – | – | – |
| Gemma 3 | **0.690** | $^{***}$0.596 | $^{***}$0.201 | $^{*}$0.642 | $^{**}$0.633 |
| InternVL3 | 0.711 | **0.711** | $^{***}$-0.231 | 0.702 | 0.664 |
| LLaVA 1.5 | **0.191** | $^{***}$0.031 | $^{***}$-0.077 | $^{***}$0.018 | $^{***}$0.043 |
| Pixtral | 0.494 | $^{*}$**0.582** | $^{***}$0.287 | 0.516 | 0.527 |
| Qwen2.5-VL | 0.628 | $^{**}$0.568 | **0.655** | $^{**}$0.520 | $^{*}$0.549 |

MT40k (Concreteness)

| | *None* | *White* | *Noise* | *ImageNet* | *Wikimedia* |
|---|---|---|---|---|---|
| | | | **All** | | |
| LLaMA 2 | 0.085 | – | – | – | – |
| Mistral | 0.638 | – | – | – | – |
| Qwen2.5 | 0.538 | – | – | – | – |
| Gemma 3 | **0.439** | 0.414 | $^{***}$-0.106 | 0.413 | 0.428 |
| InternVL3 | **0.641** | $^{***}$0.561 | $^{***}$-0.329 | 0.624 | 0.614 |
| LLaVA 1.5 | 0.047 | -0.007 | 0.058 | $^{**}$0.183 | $^{**}$**0.187** |
| Pixtral | 0.438 | $^{*}$0.365 | $^{***}$-0.021 | $^{***}$0.578 | $^{***}$**0.591** |
| Qwen2.5-VL | 0.361 | $^{***}$0.195 | $^{***}$-0.019 | $^{***}$**0.576** | $^{***}$0.562 |
| | | | **Low imagery** | | |
| LLaMA 2 | 0.178 | – | – | – | – |
| Mistral | 0.509 | – | – | – | – |
| Qwen2.5 | 0.426 | – | – | – | – |
| Gemma 3 | **0.474** | 0.448 | $^{***}$0.118 | 0.450 | 0.441 |
| InternVL3 | **0.463** | $^{***}$0.349 | $^{***}$-0.036 | $^{***}$0.298 | $^{***}$0.325 |
| LLaVA 1.5 | **0.207** | $^{***}$-0.040 | $^{**}$0.006 | 0.097 | 0.099 |
| Pixtral | **0.478** | $^{*}$0.373 | $^{***}$-0.093 | 0.430 | $^{*}$0.404 |
| Qwen2.5-VL | 0.324 | $^{***}$0.195 | $^{***}$-0.038 | **0.395** | 0.379 |
| | | | **High imagery** | | |
| LLaMA 2 | 0.001 | – | – | – | – |
| Mistral | 0.429 | – | – | – | – |
| Qwen2.5 | 0.279 | – | – | – | – |
| Gemma 3 | 0.009 | 0.031 | $^{***}$-0.230 | **0.039** | 0.022 |
| InternVL3 | 0.416 | 0.423 | $^{***}$-0.292 | **0.431** | 0.389 |
| LLaVA 1.5 | -0.082 | $^{**}$0.049 | $^{*}$0.052 | 0.041 | $^{*}$**0.077** |
| Pixtral | 0.042 | 0.065 | 0.096 | $^{***}$0.265 | $^{***}$**0.285** |
| Qwen2.5-VL | 0.161 | $^{***}$0.029 | $^{***}$-0.032 | **0.250** | 0.234 |

CP2004B (Imagery)

Table 8: Spearman correlations between model predictions and human ratings for concreteness (left) and imagery (right). *White*/*Noise* are uninformative contexts; *ImageNet*/*Wikimedia* are real-image contexts. Green/purple cells mark higher/lower within-row correlations; **boldface**/underlining mark best/worst contexts. Superscripts mark significance versus *None* ($^{*}$ $q < 0.05$, $^{**}$ $q < 0.01$, $^{***}$ $q < 0.001$; paired swap tests with Benjamini–Hochberg correction).

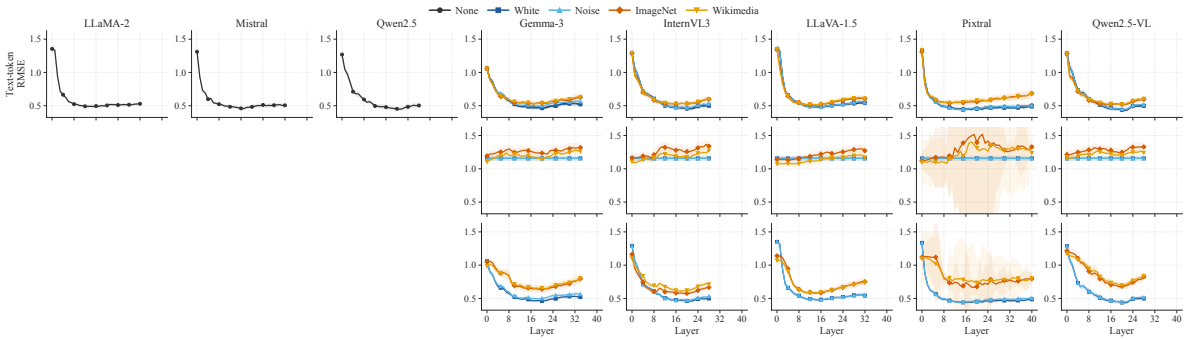
| | Win margin | | MAE reduction | |
|---|---|---|---|---|
| Variant | *ImageNet* | *Wikimedia* | *ImageNet* | *Wikimedia* |
| | | **All** | | |
| V1 | $^{***}$-23 | $^{***}$-32 | $^{***}$-0.08 | $^{***}$-0.10 |
| V2 | $^{***}$-29 | $^{***}$-37 | $^{***}$-0.12 | $^{***}$-0.13 |
| V3 | $^{***}$-20 | $^{***}$-25 | $^{***}$-0.06 | $^{***}$-0.06 |
| V4 | $^{***}$-23 | $^{***}$-29 | $^{***}$-0.06 | $^{***}$-0.08 |
| V5 | $^{*}$-6 | $^{***}$-11 | $^{***}$-0.05 | $^{***}$-0.08 |
| **Avg.** | **-20** | **-27** | **-0.07** | **-0.09** |
| | | **Abstract** | | |
| V1 | $^{***}$+44 | $^{***}$+25 | $^{***}$+0.10 | $^{***}$+0.06 |
| V2 | $^{***}$+23 | -0 | $^{***}$+0.04 | -0.00 |
| V3 | $^{***}$+19 | +7 | $^{***}$+0.03 | +0.01 |
| V4 | $^{***}$+22 | +2 | $^{***}$+0.04 | -0.00 |
| V5 | $^{***}$+84 | $^{***}$+78 | $^{***}$+0.35 | $^{***}$+0.31 |
| **Avg.** | **+38** | **+22** | **+0.11** | **+0.07** |
| | | **Concrete** | | |
| V1 | $^{***}$-64 | $^{***}$-67 | $^{***}$-0.24 | $^{***}$-0.23 |
| V2 | $^{***}$-62 | $^{***}$-62 | $^{***}$-0.26 | $^{***}$-0.24 |
| V3 | $^{***}$-46 | $^{***}$-46 | $^{***}$-0.13 | $^{***}$-0.13 |
| V4 | $^{***}$-50 | $^{***}$-49 | $^{***}$-0.15 | $^{***}$-0.14 |
| V5 | $^{***}$-81 | $^{***}$-84 | $^{***}$-0.38 | $^{***}$-0.41 |
| **Avg.** | **-61** | **-62** | **-0.23** | **-0.23** |

MT40k (Concreteness)

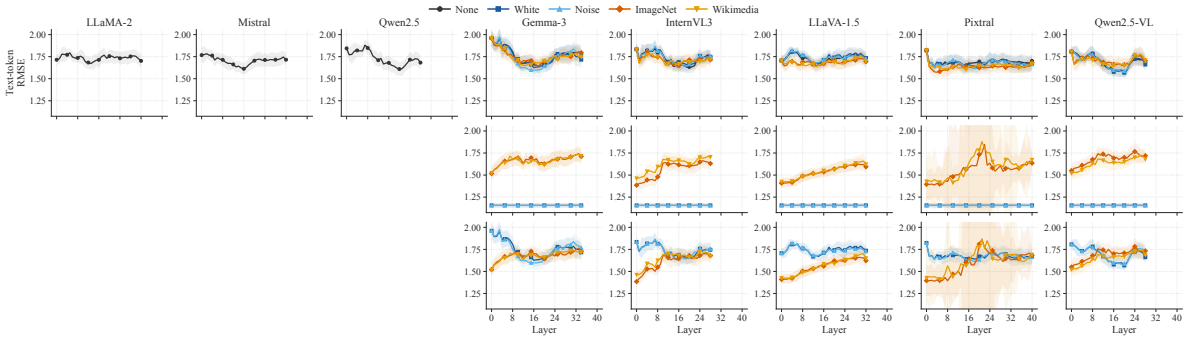
| | Win margin | | MAE reduction | |
|---|---|---|---|---|
| Variant | *ImageNet* | *Wikimedia* | *ImageNet* | *Wikimedia* |
| | | **All** | | |
| V1 | $^{***}$+40 | $^{***}$+32 | $^{***}$+0.19 | $^{***}$+0.14 |
| V2 | +0 | $^{***}$-9 | $^{**}$-0.02 | $^{***}$-0.05 |
| V3 | -4 | $^{***}$-9 | -0.01 | $^{**}$-0.02 |
| V4 | $^{***}$-28 | $^{***}$-28 | $^{***}$-0.11 | $^{***}$-0.12 |
| V5 | $^{***}$+44 | $^{***}$+38 | $^{***}$+0.22 | $^{***}$+0.18 |
| **Avg.** | **+11** | **+5** | **+0.05** | **+0.03** |
| | | **Low imagery** | | |
| V1 | $^{***}$+70 | $^{***}$+62 | $^{***}$+0.37 | $^{***}$+0.29 |
| V2 | -1 | $^{***}$-17 | $^{***}$-0.05 | $^{***}$-0.10 |
| V3 | +1 | -2 | +0.00 | -0.01 |
| V4 | $^{***}$-39 | $^{***}$-43 | $^{***}$-0.17 | $^{***}$-0.20 |
| V5 | $^{***}$+72 | $^{***}$+70 | $^{***}$+0.39 | $^{***}$+0.35 |
| **Avg.** | **+21** | **+14** | **+0.11** | **+0.06** |
| | | **High imagery** | | |
| V1 | $^{***}$+12 | +3 | $^{*}$+0.03 | -0.00 |
| V2 | +2 | -1 | +0.00 | -0.01 |
| V3 | $^{**}$-9 | $^{***}$-15 | $^{**}$-0.03 | $^{***}$-0.03 |
| V4 | $^{***}$-16 | $^{***}$-13 | $^{***}$-0.05 | $^{***}$-0.04 |
| V5 | $^{***}$+16 | +6 | $^{***}$+0.05 | $^{*}$+0.03 |
| **Avg.** | **+1** | **-4** | **-0.00** | **-0.01** |

CP2004B (Imagery)

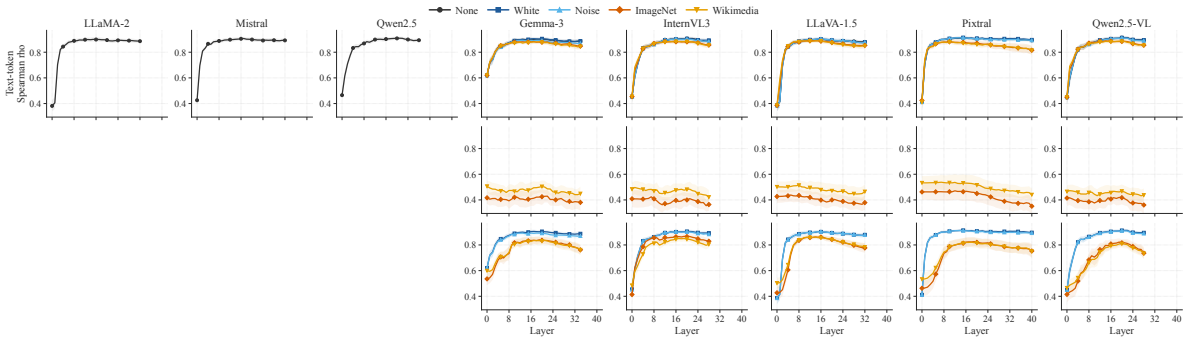
Table 9: Non-tied win margin and MAE reduction for Qwen2.5-VL mitigation prompt variants on concreteness (left) and imagery (right). *ImageNet*/*Wikimedia* are real-image contexts. Green/purple cells mark positive/negative changes, scaled within each variant and metric. Superscripts mark significance versus standard prompting after Benjamini–Hochberg correction, as in Figure 5.



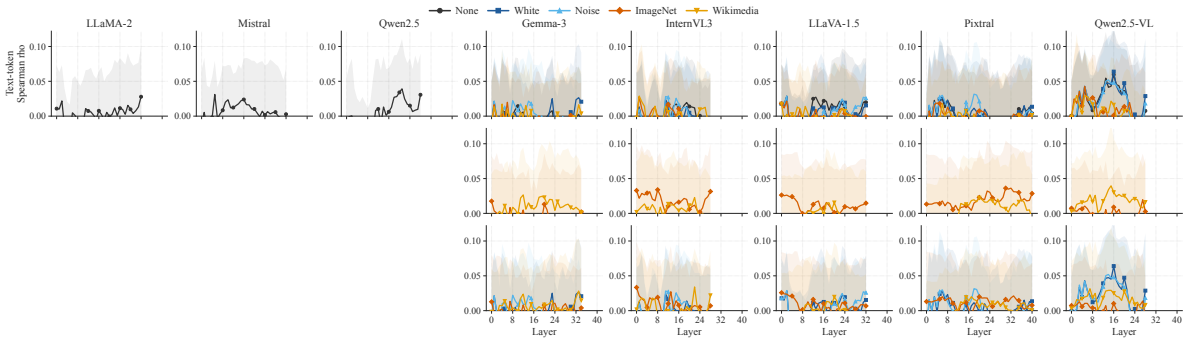

(a) RMSE




(b) RMSE, permuted labels

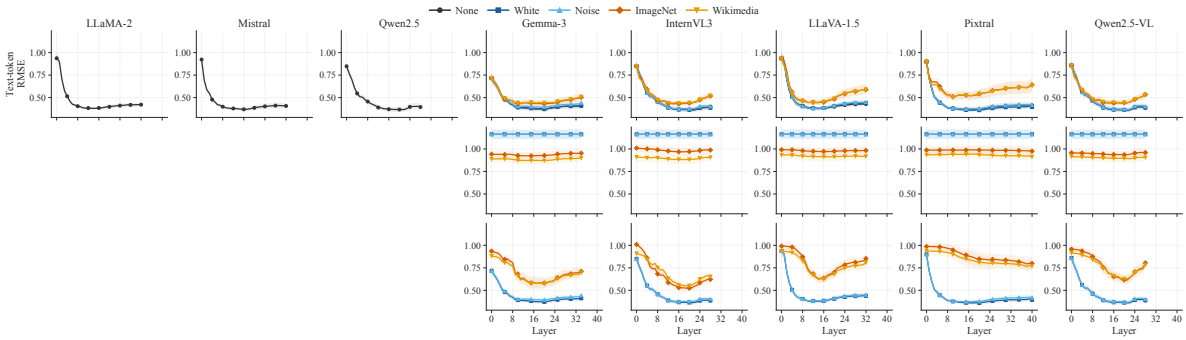



(c) Spearman

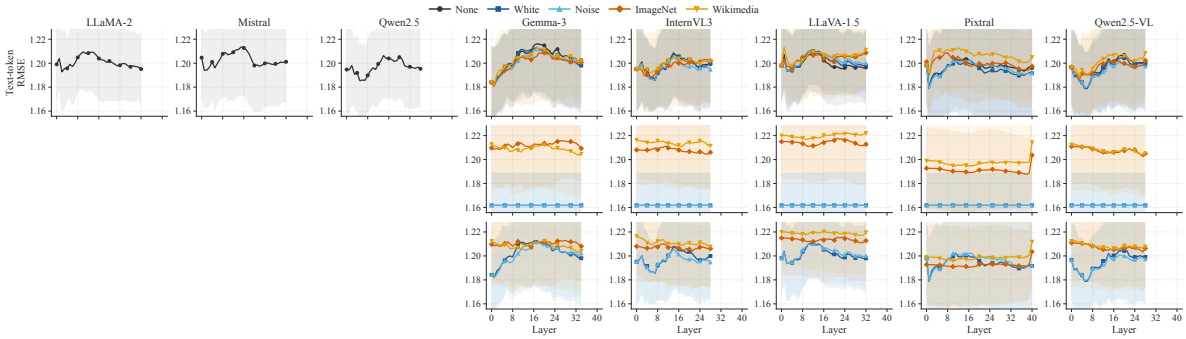



(d) Spearman, permuted labels

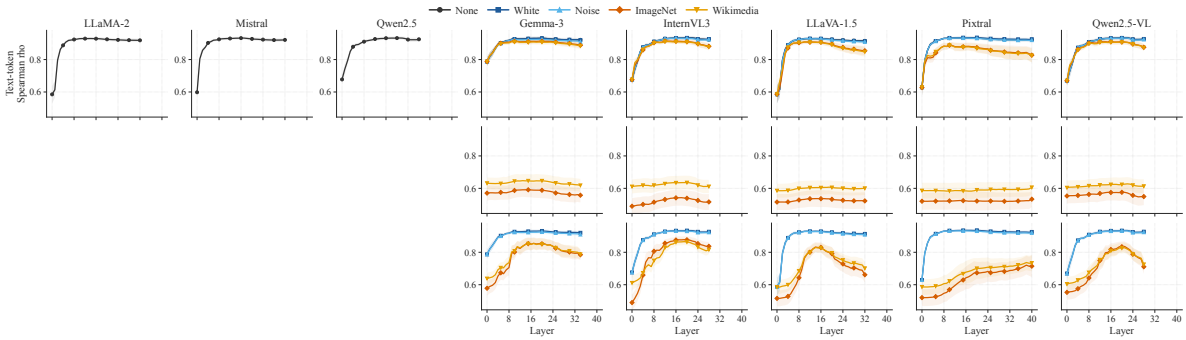
Figure 6: Layer-wise ridge probing on MT40k concreteness. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



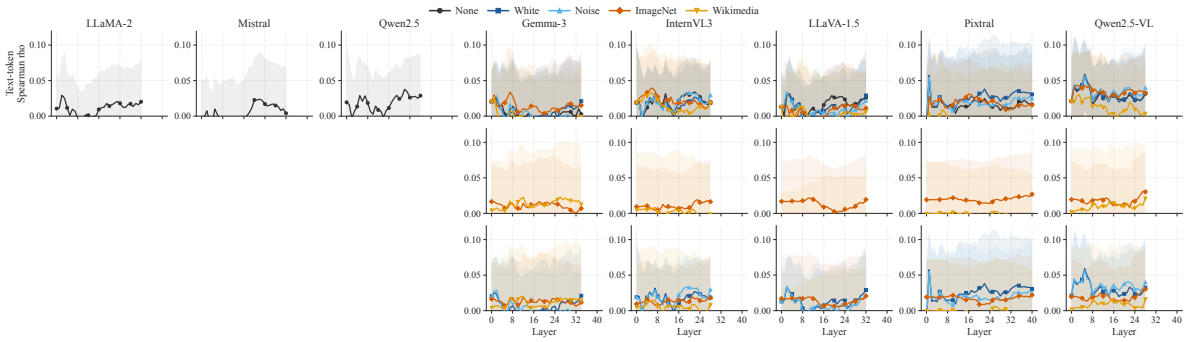

(a) RMSE




(b) RMSE, permuted labels

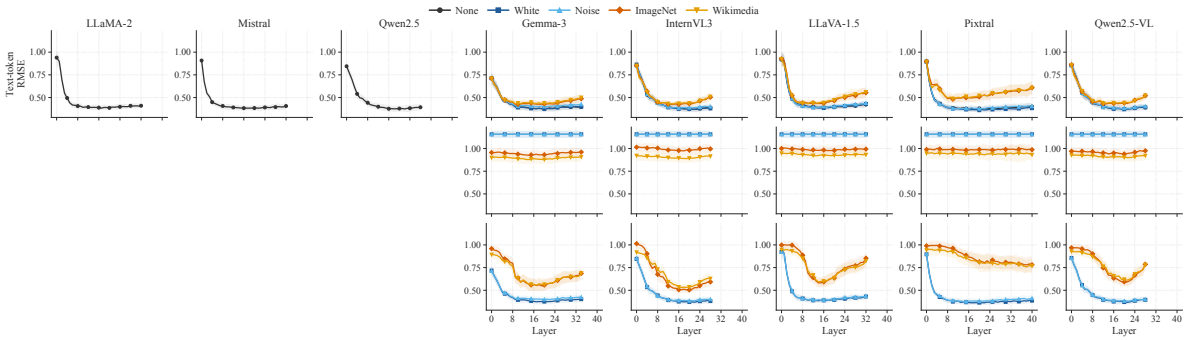



(c) Spearman

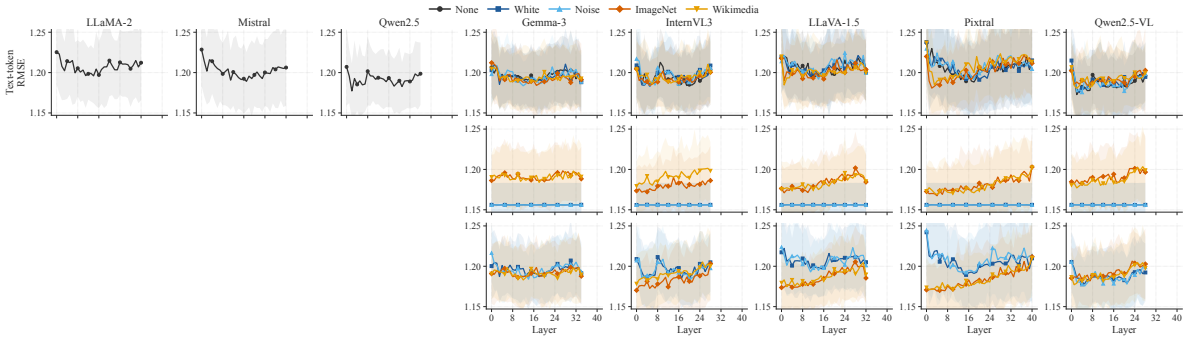



(d) Spearman, permuted labels

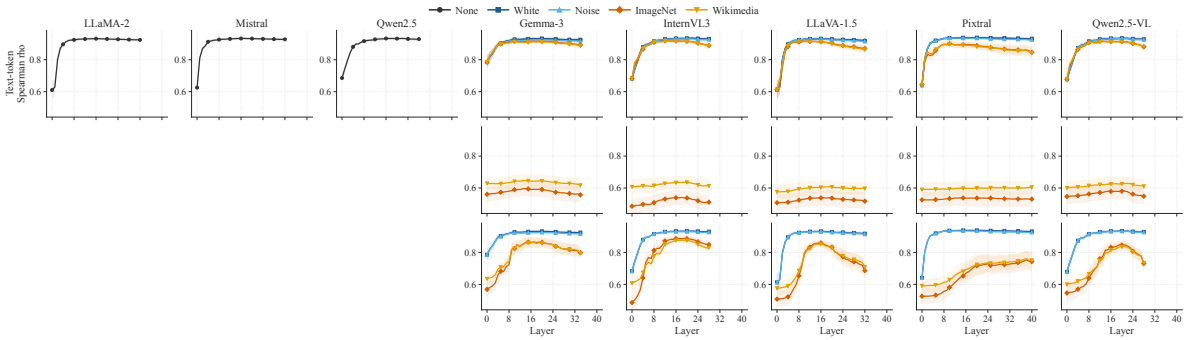
Figure 7: Layer-wise SVM probing on MT40k concreteness. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



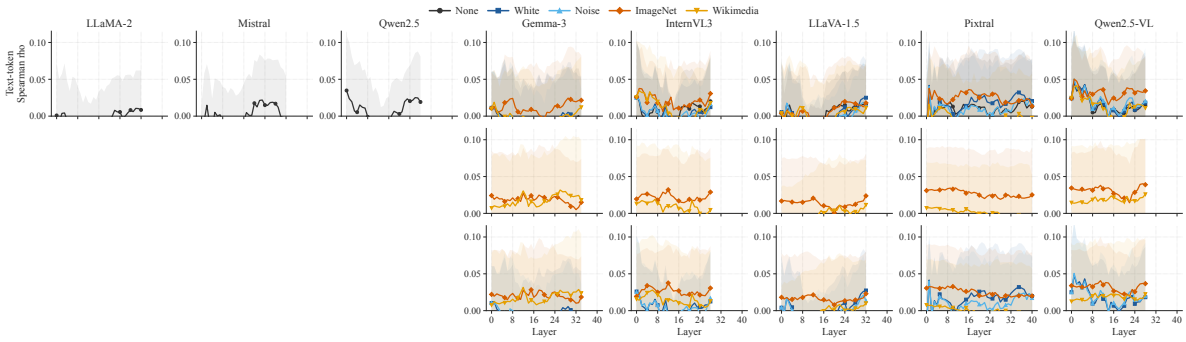

(a) RMSE




(b) RMSE, permuted labels

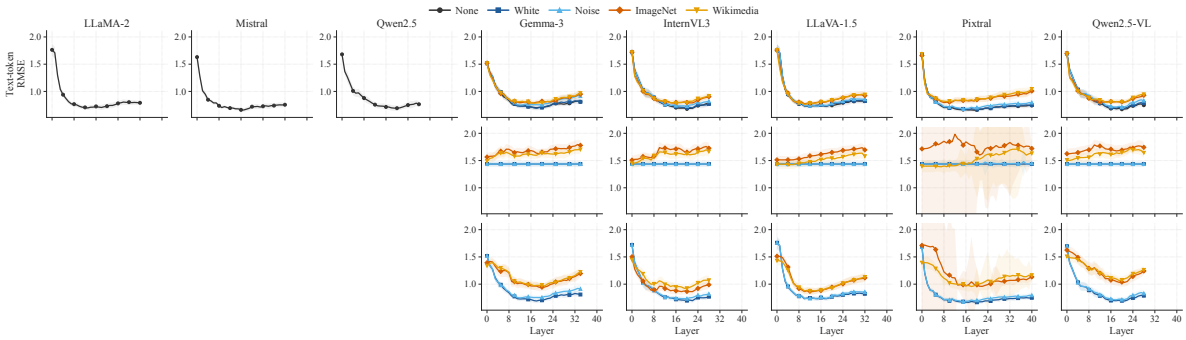



(c) Spearman

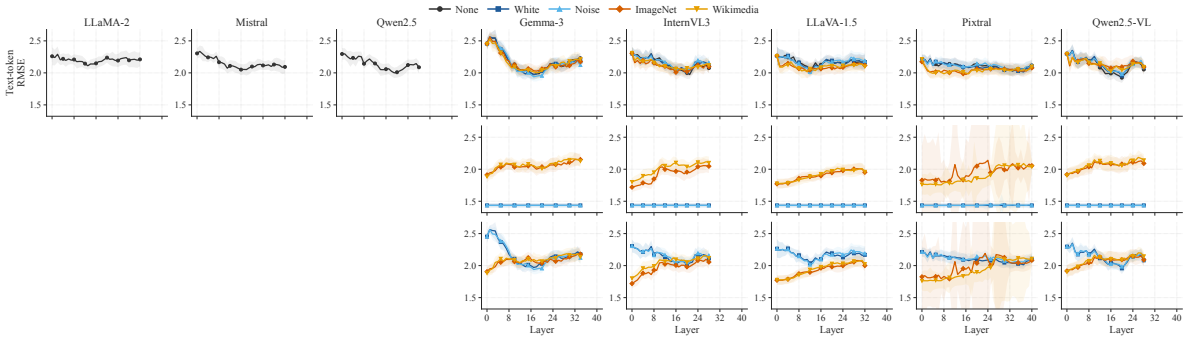



(d) Spearman, permuted labels

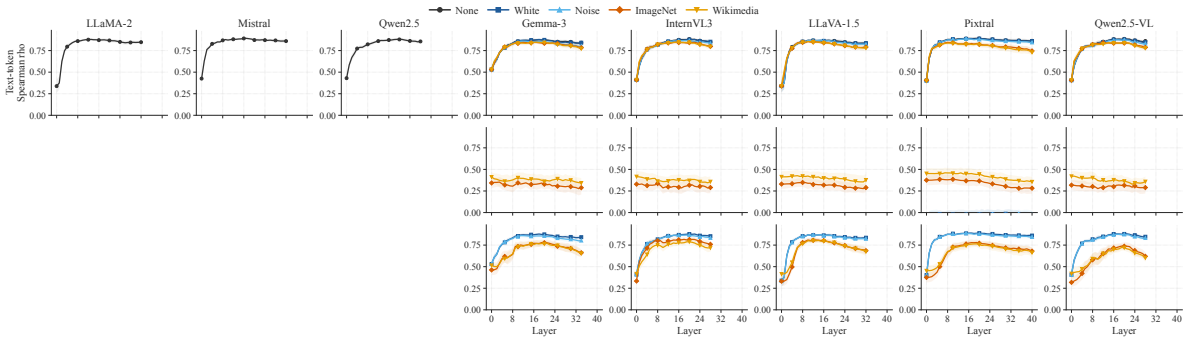
Figure 8: Layer-wise MLP probing on MT40k concreteness. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



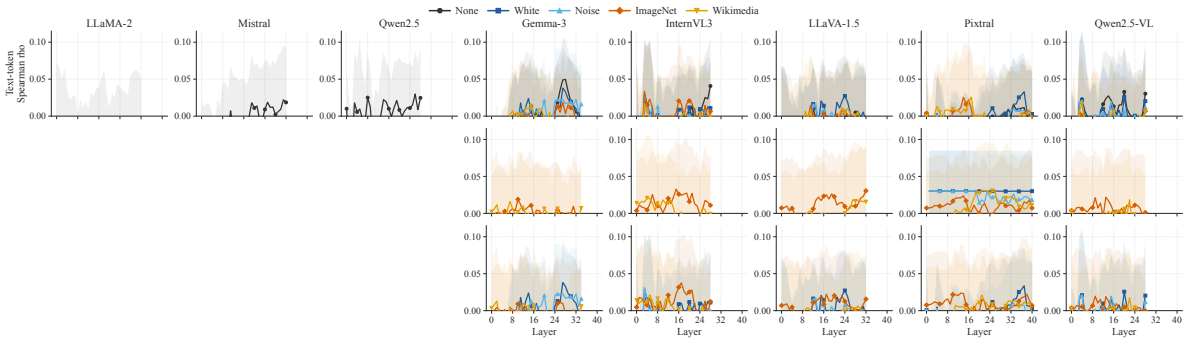

(a) RMSE




(b) RMSE, permuted labels

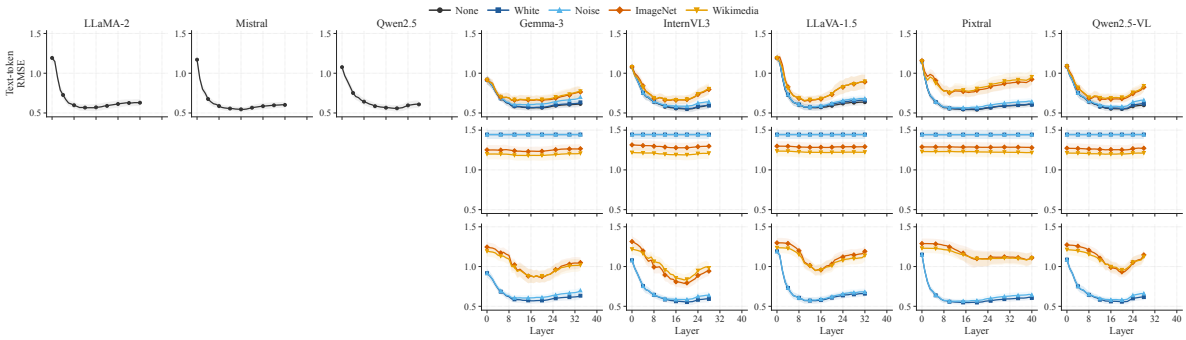



(c) Spearman

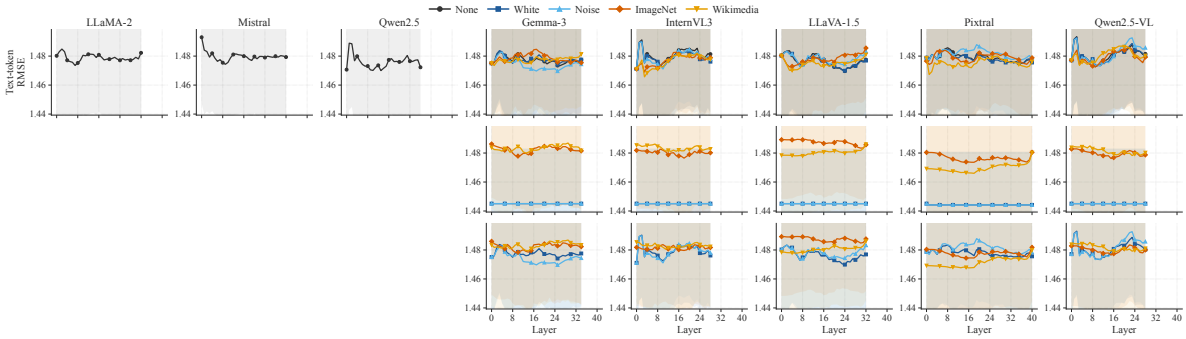



(d) Spearman, permuted labels

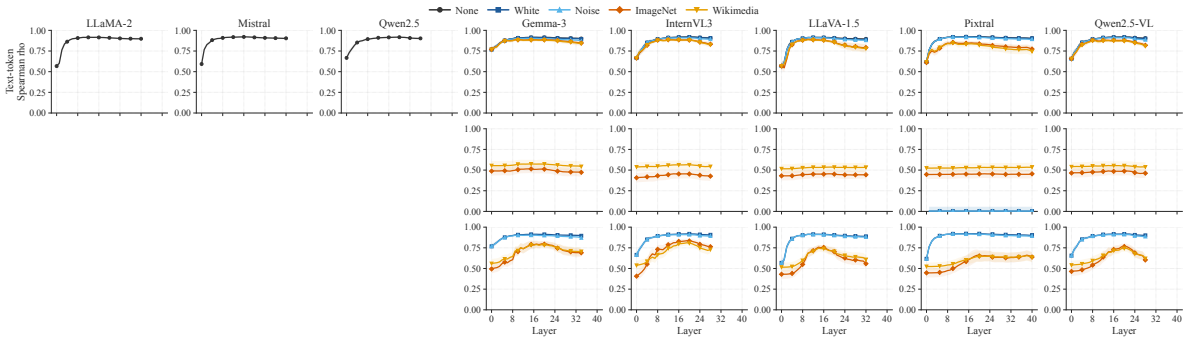
Figure 9: Layer-wise ridge probing on CP2004B imagery. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



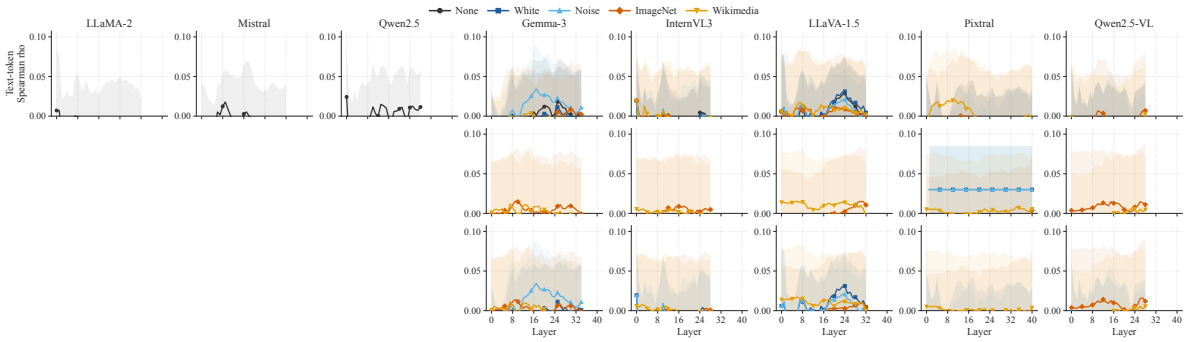

(a) RMSE




(b) RMSE, permuted labels

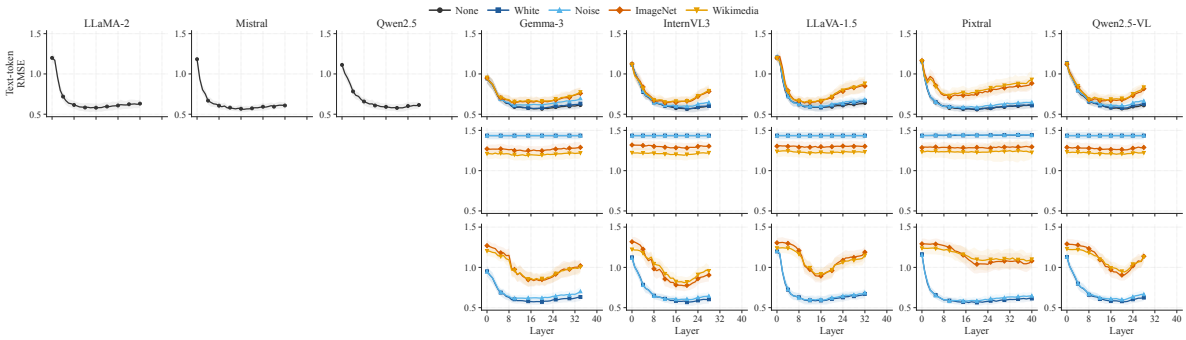



(c) Spearman

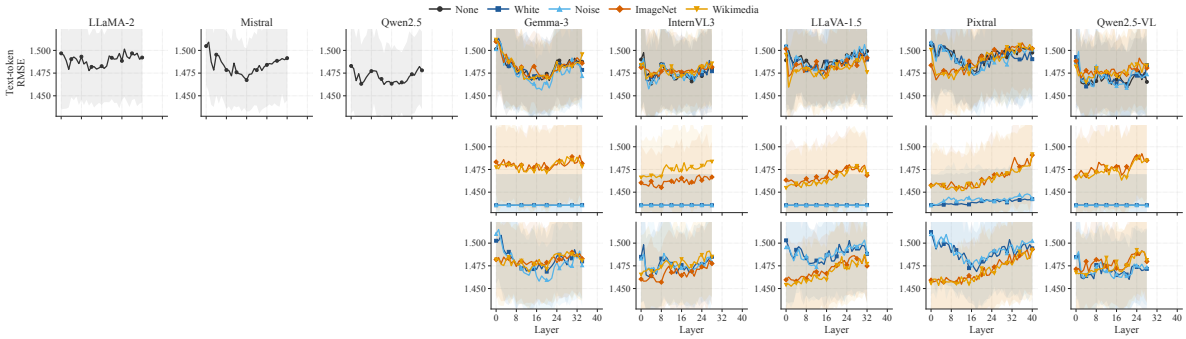



(d) Spearman, permuted labels

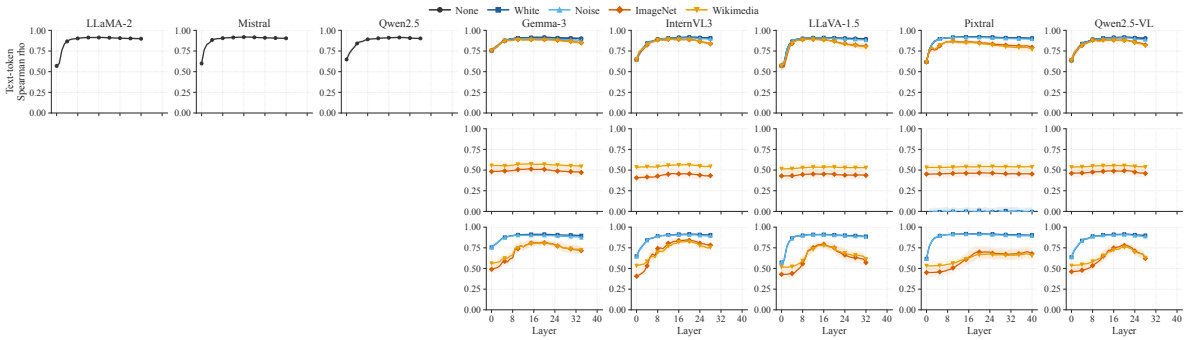
Figure 10: Layer-wise SVM probing on CP2004B imagery. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



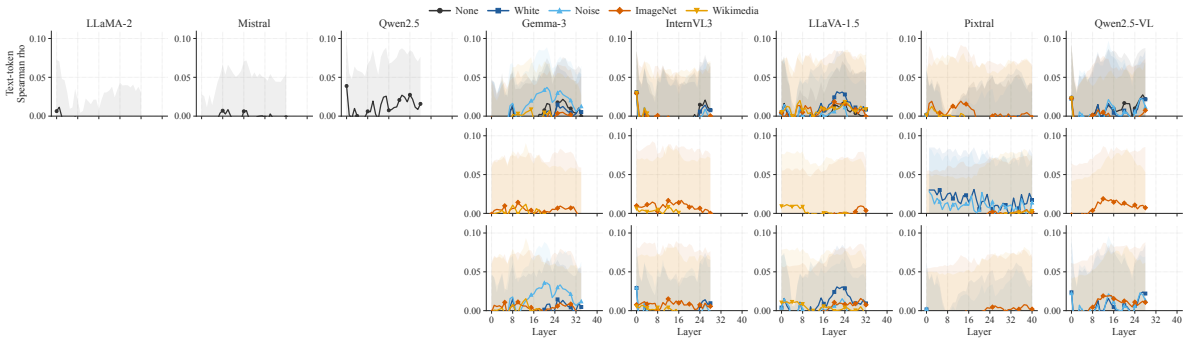

(a) RMSE




(b) RMSE, permuted labels

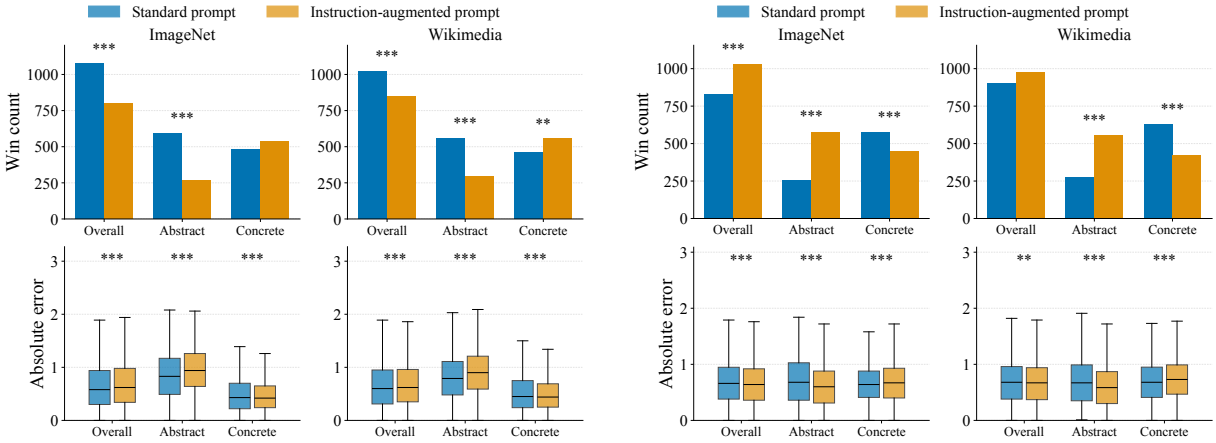



(c) Spearman




(d) Spearman, permuted labels

Figure 11: Layer-wise MLP probing on CP2004B imagery. Panels show RMSE and Spearman, each with a permuted-label baseline. VLM probes use text-token, image-token, and combined representations; bands show standard deviations across folds and prompt templates.



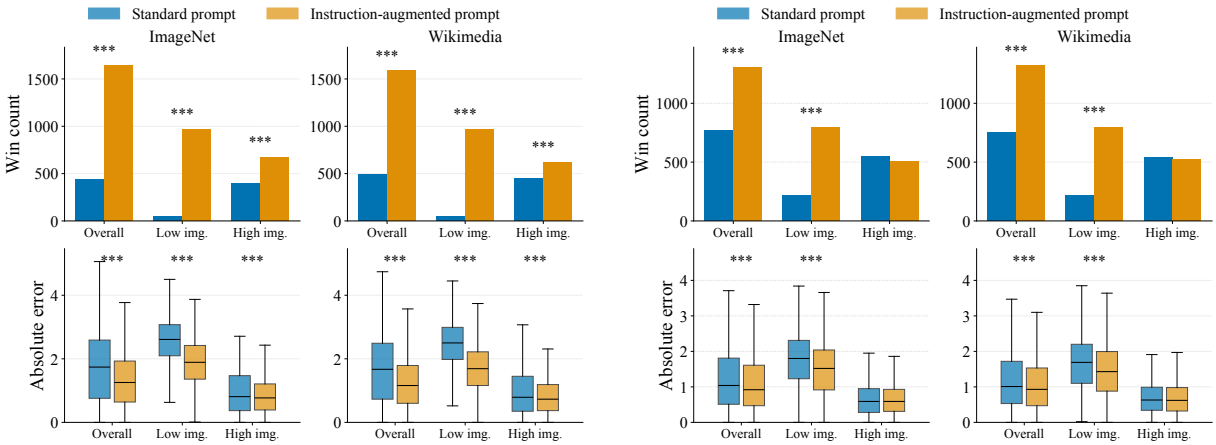

(a) Gemma 3

(b) InternVL3

(c) LLaVA 1.5

(d) Pixtral

Figure 12: Additional MT40k mitigation results for real-image contexts. Each subfigure compares standard and instruction-augmented prompting for one VLM. Top rows show non-tied win counts; bottom rows show absolute-error distributions, overall and by concreteness subset. Superscripts mark significance as in Figure 5.




(a) Gemma 3

(b) InternVL3

(c) LLaVA 1.5

(d) Pixtral

Figure 13: Additional CP2004B mitigation results for real-image contexts. Each subfigure compares standard and instruction-augmented prompting for one VLM. Top rows show non-tied win counts; bottom rows show absolute-error distributions, overall and by imagery subset. Superscripts mark significance as in Figure 5.