CAN LANGUAGE REPRESENT SEMANTIC VARIATION? A SEMANTIC BASIS FOR VLM ROBUSTNESS AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) underpin a broad range of real-world AI applications. Yet understanding their robustness under semantic variation remains difficult, since such variations are hard to formalize directly in visual space. We therefore ask whether language can provide a reliable representation of semantic variation for robustness auditing. We construct semantic directions from semantic-contrast text pairs and aggregate them across textual contexts for consistency. Cross-modal correction further calibrates these directions to the visual embedding space. Controlled visual validation shows that language-derived semantic representations align well with corresponding visual semantic variations. Building on this semantic basis, we establish a semantic robustness auditing framework for VLMs that quantifies prediction stability along semantic variation with closed-form boundary margins. The framework supports auditing across data, prompt, and model contexts, while enabling fine-grained diagnosis of where robustness is preserved or degraded. Our audits reveal systematic variation in semantic robustness across contexts alongside stable semantic patterns. Domain-specific applications further show that the audit localizes fragile patterns aligned with task priors and guides model selection under target-domain robustness requirements. Overall, our work provides a language-derived semantic basis for auditing and diagnosing VLM robustness under semantic variation.

1 INTRODUCTION

Vision-Language Models (VLMs) have shown strong transferability across recognition and multimodal understanding tasks by aligning images and texts in a shared embedding space (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2022; Zou et al., 2023). As these models are increasingly used in real-world settings, understanding whether their predictions remain stable under semantic variation becomes increasingly important, yet such robustness remains poorly understood. While existing studies have examined out-of-distribution generalization, adversarial vulnerability, and prompt adaptation in VLMs (Ming et al., 2022; Zhao et al., 2023; Schlarmann et al., 2024; Zhou et al., 2022b;a), and recent work has also explored concept-based interpretation and holistic evaluation (Bhalla et al., 2024; Kim et al., 2023; Lee et al., 2024; Liang et al., 2024; Xu et al., 2024), the fundamental lens remains one of overall performance and broad capability assessment. This leaves semantic robustness under data, prompt, and model shifts largely uncharacterized, even though these shifts are central to understanding the reliability of VLMs in real-world use. This gap becomes sharper in practical audits, where semantic counterfactuals and fine-grained labels are rarely available.

To resolve these fundamental limitations, this work presents a semantic robustness auditing framework for VLMs. VLMs combine a natural-language interface with similarity-based prediction in a shared embedding space, allowing language to provide an explicit representation of open-vocabulary semantic variation. This representation enables semantic robustness to be audited without manually annotated image edits. For a target semantic attribute, we instantiate a base prompt and its semantic counterpart, where the counterpart introduces the target attribute while preserving the template and category context. A practical audit remains nontrivial, because the resulting directions can vary with prompt wording, and their transfer to the visual side is further affected by residual cross-modal mismatch. We address these challenges by aggregating semantic directions over textual contexts to reduce template and category artifacts (see Fig. 1(a)), and by learning a cross-modal correction to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

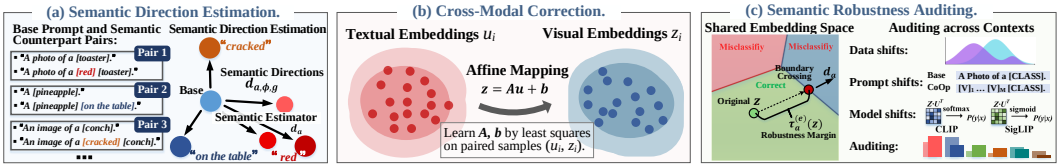


Figure 1: Overview of semantic robustness auditing framework. The framework consists of three components: (a) semantic direction estimation from prompt pairs, (b) cross-modal correction for transferring textual directions to the visual side, and (c) semantic robustness auditing via boundary margins. Overall, these components support unified auditing across data, prompts, and VLMs.

transfer the text-defined directions to the visual side of the shared space (see Fig. 1(b)). We show that the language-derived representations align well with corresponding visual semantic variations.

On this semantic basis, we derive a closed-form semantic boundary margin for each sample under a semantic direction, defined by the first decision-boundary crossing along the audited semantic direction. Semantic robustness can then be analyzed through statistics of the resulting margins, rather than through overall accuracy. This yields a unified audit framework for comparing semantic robustness across evaluation contexts, including data, prompt, and model shifts. Overall, our framework audits semantic robustness via the first decision-boundary crossing along each audited semantic direction, yielding comparable margins across evaluation contexts, as shown in Fig. 1(c).

Experiments first validate that the language-derived semantic representations align well with controlled visual semantic variation. Our audit framework then supports two complementary forms of robustness analysis. At the context level, it reveals clear and structured variation in semantic robustness across evaluation contexts that is not captured by standard accuracy evaluation. Across datasets, prompt learners, and VLM families, it further uncovers persistent attribute-level patterns, class-level robustness disparities, and structured confusion behavior. Beyond such overall comparisons, it also supports fine-grained diagnosis of robustness degradation under prompt and model scaling, revealing that robustness gains are often highly uneven. Applications on GTSRB (Stallkamp et al., 2011) and EuroSAT (Helber et al., 2019) further show that the framework localizes domain-specific fragile patterns and guides target-domain model selection. Overall, these results position our semantic robustness auditing framework as a principled and interpretable tool for understanding the reliability of VLMs. The main contributions of this work are summarized as follows:

1. We establish language as a basis of semantic variation through context-consistent directions and cross-modal calibration, yielding alignment with corresponding visual variation.
2. We formulate a semantic robustness auditing framework with closed-form semantic boundary margins, supporting auditing across data, prompt, and model contexts.
3. Our audits uncover structured differences in semantic robustness across data, prompt, and model contexts, including non-uniform scaling across classes and domain-dependent model preferences.

2 Related Work

Robustness in VLMs. VLMs, such as CLIP (Radford et al., 2021), are built on a shared image-text embedding space, which substantially improved zero-shot transfer across tasks (Li et al., 2022; Alayrac et al., 2022; Zou et al., 2023). Despite their strong capabilities, VLMs still exhibit clear fragility under distribution shifts, adversarial perturbations, and prompt variations (Ming et al., 2022; Shu et al., 2023; Schlarmann et al., 2024). For instance, prior work has shown that VLMs are sensitive to out-of-distribution data (Ming et al., 2022), which has motivated methods for distribution adaptation and robust generalization (Shu et al., 2023; Cho et al., 2023). Related studies also reveal adversarial vulnerability in large VLMs (Zhao et al., 2023), while Robust CLIP improves robustness through adversarial fine-tuning of the CLIP vision encoder (Schlarmann et al., 2024). In parallel, prompt learning has become a main route for adapting VLMs. CoOp (Zhou et al., 2022b) learns continuous context vectors for downstream classes, while CoCoOp (Zhou et al., 2022a) conditions prompts on individual instances. At the model level, robustness has also been examined across a series of CLIP variants and alignment improvements, including OpenCLIP (Cherti et al., 2023), EVA-CLIP (Sun et al., 2023), SigLIP (Zhai et al., 2023), and AlignCLIP (Eslami & de Melo, 2024).

VLM Interpretation and Evaluation. As the vulnerabilities of VLMs become increasingly evident, recent work has also focused on interpreting their behavior and evaluating their limitations.

SpLiCE represents CLIP image embeddings as sparse combinations of human-readable text concepts (Bhalla et al., 2024). CounTEX defines concepts directly from text and produces counterfactual concept explanations by aligning classifier features with CLIP space (Kim et al., 2023). Other work further studies VLM representations through preferred concept descriptions and analyses of CLIP explainability (Esfandiarpoor et al., 2024; Li et al., 2025). Recent benchmarks have moved toward more holistic evaluation. VHELM (Lee et al., 2024) evaluates VLMs across a broad range of capability and risk dimensions under a unified evaluation framework. Xu et al. (2024) build a comprehensive evaluation hub for large VLMs by combining large-scale benchmark testing with human preference comparison. More targeted benchmarks examine specific weaknesses such as color understanding and compositional reasoning (Liang et al., 2025; Huang et al., 2024).

3 Problem Statement

We consider VLMs (Radford et al., 2021; Zhai et al., 2023) that encode images and texts into a shared normalized embedding space $\mathbb{S}^{d-1} := \{e \in \mathbb{R}^d : \|e\|_2 = 1\}$. For an image x and a prompt set $\{t_c : c \in \mathcal{C}\}$ over a class set $\mathcal{C}$, the visual encoder f_{img} and textual encoder f_{text} map both modalities into $\mathbb{S}^{d-1}$. We denote the image and text embeddings by $z := f_{\text{img}}(x)$ and $u_c := f_{\text{text}}(t_c)$. The model then predicts by selecting the class embedding most similar to z as $f(z) := \arg\max_{c \in \mathcal{C}} \langle z, u_c \rangle$.

Our Objective. We aim to audit how semantic robustness changes across evaluation contexts, with each context defined by the data distribution, the prompt setting, and the underlying VLM. Given a semantic attribute a, our goal is to define a measure that quantifies the extent over which a prediction remains invariant under semantic variation along a, and then compare this quantity across different data distributions, models, and prompt settings.

Semantic Auditing Model. For a semantic attribute a and an image embedding z, we formalize controlled semantic variation in the embedding space by a transformation

$$\gamma_a(\delta; z) : [0, \delta_{\max}] \to \mathbb{S}^{d-1}, \tag{1}$$

parameterized by an extent $\delta \in [0, \delta_{\max}]$, with $\gamma_a(0; z) = z$. $\gamma_a(\delta; z)$ represents moving z along attribute a, with larger δ indicating a greater semantic variation. Let $\hat{y} := f(z)$ denote the prediction on x. The prediction is semantically robust along attribute a up to extent δ if it remains invariant over the segment $[0, \delta]$. Accordingly, we define the semantic boundary margin as

$$\tau_a(x) := \sup\Big\{\delta \in [0, \delta_{\max}] : f(\gamma_a(r; z)) = \hat{y},\ \forall r \in [0, \delta]\Big\}. \tag{2}$$

For each semantic attribute a, $\tau_a(x)$ quantifies the local semantic robustness of the prediction on x under controlled semantic variation. At the population level, semantic robustness is characterized by the distribution of $\tau_a(x)$. Semantic robustness change under data, model, or prompt variation is then measured by changes in the corresponding boundary-margin statistics across evaluation contexts.

4 Method

In this section, we present our framework for semantic robustness auditing in VLMs. We begin by constructing attribute directions from textual prompts to model semantic variation in the shared embedding space, and then transfer these text-defined directions to the visual side through cross-modal correction, yielding a reliable auditing model. Building on this model, we quantify prediction stability under semantic variation and derive auditing statistics across evaluation contexts.

4.1 Semantic Direction Estimation

A practical instance of the semantic auditing model $\gamma_a(\delta; z)$ in Eq. 1 requires, for each semantic attribute a, a representation that specifies the intended variation in semantics along a. We construct this representation from text prompts, enabled by contrastive vision-language pre-training that brings matched image-text pairs close and separates mismatched pairs (Radford et al., 2021; Jia et al., 2021). As a result, textual embeddings can provide a structured semantic reference in the shared embedding space, making language a natural way to specify open-vocabulary semantic attributes. Image embeddings, by comparison, reflect the combined context of the whole image and therefore

usually mix multiple semantic attributes. This motivates constructing attribute representations from paired textual prompts, where a base prompt fixes the context and its semantic counterpart changes only the target attribute. For attribute a, template $\phi \in \Phi$, and class domain $g \in \mathcal{G}$, let $t_{\phi,g}$ denote the base prompt and $t'_{a,\phi,g}$ its semantic counterpart that differs from $t_{\phi,g}$ only in realizing a. The corresponding text embeddings and semantic direction are then given by

$$u_{\phi,g} := f_{\text{text}}(t_{\phi,g}), \quad u'_{a,\phi,g} := f_{\text{text}}(t'_{a,\phi,g}), \quad d_{a,\phi,g} := \text{norm}(u'_{a,\phi,g} - u_{\phi,g}). \tag{3}$$

Here, $d_{a,\phi,g}$ denotes the normalized semantic direction induced by attribute a, and $\text{norm}(v) := v/\|v\|_2$. It thus defines a transformation on image embedding z as $\gamma_a(\delta; z) := \text{norm}(z + \delta d_{a,\phi,g})$.

The construction above turns a target attribute into an explicit semantic transformation in the shared embedding space, yielding a concrete auditing model. Its reliability first requires that directions instantiated under different templates and class domains share a common attribute component rather than prompt or class artifacts. Once such a shared component exists, the remaining variation should be traced back to the template ϕ and class domain g, which instantiate the textual context in Eq. 3. We examine these two issues next as a reliability analysis of the resulting auditing model.

Question 1: Does a shared attribute direction exist across textual contexts? Although VLMs induce a structured semantic geometry in the shared embedding space, it remains unclear whether the direction associated with a semantic attribute is preserved across different contexts. To examine this, we analyze pairwise alignment among context-specific directions and summarize their similarity distributions across attribute groups. For two attributes a_1 and a_2, the heatmap entry in Fig. 2 is $\mathbb{E}_{\xi,\xi' \sim \Xi}[\langle d_{a_1,\xi}, d_{a_2,\xi'} \rangle]$, where $\Xi := \Phi \times \mathcal{G}$ and $\mathcal{G}$ is instantiated by ImageNet-1K classes (Deng et al., 2009). This averages directional alignment across templates and categories. Strong diagonal entries show high within-attribute consistency across contexts, while smaller off-diagonal entries indicate cross-attribute separation. For the density plot, attributes that vary within the same attribute family (e.g., color or texture) are treated as one attribute family. The plot shows that within-attribute similarities concentrate at high values, while within-family similarities shift lower and across-family similarities concentrate near zero. These results indicate that each attribute admits a shared direction in the embedding space, and that $d_{a,\phi,g}$ should be understood as a context-dependent estimate rather than a prompt-specific artifact.

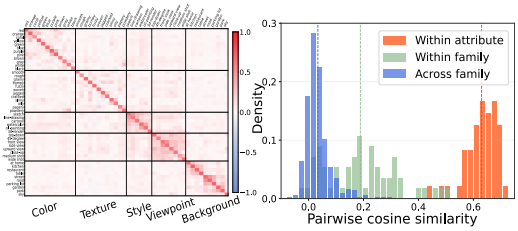
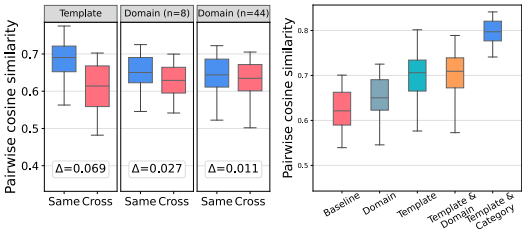


Figure 2: Pairwise alignment of semantic directions on ImageNet. The heatmap shows average cosine similarities over ImageNet-1K classes and prompt templates. The density plot summarizes the same similarities under an attribute-family grouping, showing high within-attribute consistency and lower cross-attribute alignment.

Question 2: What drives residual variation in the direction estimate? The consistency analysis above reveals a shared semantic direction, but the estimates $d_{a,\phi,g}$ still vary across textual contexts. Since each estimate is instantiated by a template ϕ and a class domain g, we next examine how these two indices account for the remaining variation. To quantify their effects, Figure 3(a) compares the average alignment under a grouping rule $\mathcal{G}$, defined for same-group pairs as $\mu_a^{\text{same}}(\mathcal{G}) := \mathbb{E}[\langle d_{a,\xi}, d_{a,\xi'} \rangle \mid \pi_{\mathcal{G}}(\xi) = \pi_{\mathcal{G}}(\xi')]$, with its cross-group counterpart defined by conditioning on $\pi_{\mathcal{G}}(\xi) \neq \pi_{\mathcal{G}}(\xi')$. We consider three rules, templates, coarse class domains derived from WordNet (Miller, 1995), and fine-grained class domains obtained with GPT-5.4 (Achiam et al., 2023). The results show that template grouping yields the largest gap, indicating that prompt wording accounts for much of the remaining variation, while both coarse and fine class-domain groupings have smaller effects. Together with the analysis above, this motivates aggregation over multiple templates and class domains to reduce residual contextual effects.



(a) Similarity across groups. (b) Similarity of aggregation.

Figure 3: Comparison of semantic direction consistency under different grouping rules on ImageNet. (a) Same-vs-cross pairwise cosine similarity under template, coarse-domain, and fine-domain groupings. (b) Within-attribute cosine similarity under different estimators, including a single-prompt baseline and aggregated directions over templates and class domains.

Final estimator. The two analyses above point to a clear estimator. For each attribute a, we aggregate semantic directions over a selected set of textual contexts so as to preserve the shared semantic component while averaging out context-specific variation. Concretely, for $\Xi_a \subseteq \Phi \times \mathcal{G}$, we first compute $d_a^\star := \frac{1}{|\Xi_a|} \sum_{(\phi,g)\in\Xi_a} d_{a,\phi,g}$ and then set $d_a := \mathrm{norm}(d_a^\star)$. To formalize this construction, we write the contextual direction as

$$d_{a,\phi,g} = m_a + b_{a,\phi}^{\mathrm{tmp}} + b_{a,g}^{\mathrm{cls}} + \varepsilon_{a,\phi,g}, \tag{4}$$

where m_a denotes the shared semantic direction, $b_{a,\phi}^{\mathrm{tmp}}$ and $b_{a,g}^{\mathrm{cls}}$ capture the deviations induced by template ϕ and class domain g, respectively, and $\varepsilon_{a,\phi,g}$ collects the remaining residual.

Proposition 4.1. *Under the decomposition in Eq. 4, suppose the three perturbation families are centered, mutually uncorrelated, and uncorrelated across distinct indices within each family, with* $\mathbb{E}\|b_{a,\phi}^{\mathrm{tmp}}\|_2^2 \le \sigma_{\mathrm{tmp}}^2$, $\mathbb{E}\|b_{a,g}^{\mathrm{cls}}\|_2^2 \le \sigma_{\mathrm{cls}}^2$, *and* $\mathbb{E}\|\varepsilon_{a,\phi,g}\|_2^2 \le \sigma_\varepsilon^2$. *Let* $\Xi_a = \Phi_a \times \mathcal{G}_a$ *with* $n_\phi := |\Phi_a|$ *and* $n_g := |\mathcal{G}_a|$, *and define* $d_a := \mathrm{norm}(\frac{1}{n_\phi n_g} \sum_{\phi\in\Phi_a} \sum_{g\in\mathcal{G}_a} d_{a,\phi,g})$. *Since* $\|m_a\|_2 = 1$, *we have*

$$\mathbb{E}\big[1 - \langle d_a, m_a\rangle\big] \le 2\left(\sigma_{\mathrm{tmp}}^2/n_\phi + \sigma_{\mathrm{cls}}^2/n_g + \sigma_\varepsilon^2/(n_\phi n_g)\right). \tag{5}$$

The proof is provided in Appx. A.1. Proposition 4.1 makes the roles of the two contextual factors explicit. Averaging over more templates reduces template-induced variation, while averaging over more class domains reduces domain-induced variation. Combining both yields the most stable estimate of the shared semantic direction.

Figure 3(b) offers empirical support for this estimator. Compared with the single-template generic-object baseline (Kim et al., 2023) (e.g., "*a photo of a [attribute] object*"), all aggregated estimators substantially improve within-attribute direction consistency. Here, categories refer to the ImageNet-1K classes used to instantiate the paired prompts. Aggregating over templates yields the largest gain over the baseline, while further aggregating over categories brings an additional improvement. The best consistency is achieved by aggregating directions computed from all templates and categories for each attribute. We thus adopt this joint aggregation scheme for semantic direction construction.

4.2 Cross-Modal Correction

The direction constructed in Sec. 4.1 specifies the semantic variation to be audited. However, the audit is performed on visual embeddings, where residual modality gaps remain in the shared embedding space (Liang et al., 2022). Therefore, directly using this direction entangles the target semantics with modality mismatch. To obtain a reliable semantic auditing model, we introduce a cross-modal correction that calibrates the estimated semantic direction for visual embeddings.

Assessing Correction Complexity. Let $\{(u_i, z_i)\}_{i=1}^N$ be paired text and image embeddings, where each pair is assumed to be semantically aligned. We assume a mapping $\psi : \mathbb{R}^d \to \mathbb{R}^d$ from text embeddings to image embeddings such that $z = \psi(u)$, and define the residual correction by $r(u) := \psi(u) - u$. Assuming r is twice differentiable in a neighborhood of a reference point u_0, a second-order expansion of r around u_0 yields an equivalent expansion of ψ of the form

$$\hat{z}_i = Au_i + b + \frac{1}{2}\mathcal{H}[(u_i - u_0), (u_i - u_0)] + R(u_i; u_0), \tag{6}$$

where $Au_i + b$ captures the affine component, $\mathcal{H}$ represents quadratic bilinear form, and $R(u_i; u_0)$ collects the remaining higher-order residual. This expansion motivates a sequence of increasingly expressive mappings: no correction, translation, affine, and quadratic. We fit the corresponding parameters by least squares on paired embeddings and assess how much mapping complexity is needed to capture the residual cross-modal mismatch.

To assess how well each mapping captures the cross-modal correspondence, we use three complementary metrics. The first is the median paired cross-modal similarity, defined as $s_{\mathrm{pair}} := \mathrm{median}_i\langle \hat{z}_i, z_i\rangle$, which directly measures how well the corrected text embedding matches its paired image embedding. While it implicitly assumes that each text-image pair is exactly matched, this assumption does not always hold in evaluation. To reduce reliance on this assumption, we further consider two neighborhood-based metrics that assess whether the semantic structure is preserved across modalities. Specifically, we compute the rank correlation between the text-side and image-side neighborhood orderings, and the top-10 neighbor overlap between the corresponding nearest-neighbor sets. Figure 4 shows a consistent pattern across all three metrics on ImageNet, using

each image and its corresponding label text as the image-text pair. The translation model yields only modest improvement over the raw baseline, which is similar in spirit to prior centering-based heuristics (Bhalla et al., 2024). The affine model captures most of the residual modality gap across all three metrics, while the quadratic extension provides only marginal additional improvement. This indicates that the dominant cross-modal mismatch is already well captured by an affine mapping.

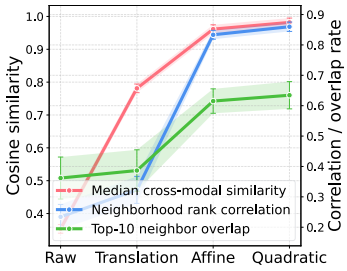


Figure 4: Effect of cross-modal correction complexity. Results are reported for three metrics, showing that the affine correction captures most of the residual modality gap, with only marginal additional gains from the quadratic correction.

Cross-Modal Correction. Based on the analysis above, we instantiate cross-modal correction with the affine map $\psi(u) := Au + b$, where (A, b) are fitted by least squares on paired text and image embeddings. For each contextual prompt pair, we map and normalize the paired text embeddings before computing the semantic direction, and aggregate the resulting directions to obtain $\tilde{d}_a$. The semantic auditing path is then instantiated as $\gamma_a(\delta; z) := \mathrm{norm}(z + \delta\tilde{d}_a)$. In this way, the text-defined semantic direction is calibrated to the image embedding space, yielding a reliable semantic auditing model.

4.3 Semantic Robustness Auditing

With the established semantic auditing model, we next audit semantic robustness of VLMs. Given a constructed semantic direction, rather than treating robustness as an informal notion, we rigorously measure semantic robustness by the margin of first decision-boundary crossing along this direction. Since the same attribute can have different semantic scales under different prompt designs and VLMs, we normalize the margins by a context-specific scale factor for cross-context auditing.

Boundary Margin and Auditing Context. For an image-text pair (x, t_y), write $\hat{y} = f(z)$ for the predicted class. Let $\{u_c\}_{c\in\mathcal{C}}$ be the class embeddings and $\tilde{d}_a$ the corrected direction for attribute a. Under the nearest prompt decision rule of VLMs, the first flip from $\hat{y}$ to a competing class occurs when the transformed image embedding reaches a pairwise decision boundary, which gives a closed-form margin along the audited direction. For each competing class $c \neq \hat{y}$, define $\alpha_{\hat{y},c} := \langle z,\, u_{\hat{y}} - u_c\rangle$ and $\beta_{\hat{y},c} := \langle \tilde{d}_a,\, u_{\hat{y}} - u_c\rangle$. The pairwise decision margin along the semantic transformation is

$$m_{\hat{y},c}(\delta; x) := \big\langle \gamma_a(\delta; z),\, u_{\hat{y}} - u_c \big\rangle = (\alpha_{\hat{y},c} + \delta\, \beta_{\hat{y},c})/\|z + \delta\tilde{d}_a\|_2. \tag{7}$$

Hence, the first boundary crossing along the audited direction gives the robustness margin as

$$\tau_a(x) = \min\{\delta_{\max}, \min_{c\neq\hat{y}:\beta_{\hat{y},c}<0} (-\alpha_{\hat{y},c}/\beta_{\hat{y},c})\}. \tag{8}$$

We define an auditing context e by the audited data distribution, the prompt design, and the VLM. The prompt design and VLM determine the classifier and semantic auditing model, while the audited data distribution specifies the samples over which the margins are aggregated. We write $\tau_a^{(e)}(x)$ for the sample-level semantic robustness margin under the auditing model associated with context e.

Semantic Scale Normalization. Although $\tau_a^{(e)}(x)$ gives the first boundary crossing under a fixed auditing context, its raw value depends on the semantic scale of the context-specific audited axis and is not always directly comparable across contexts. Changing the prompt design or VLM can rescale the semantic axis and therefore affect the magnitude of the raw margin. To reduce this scale effect, we normalize the raw margin by a context-specific factor. Concretely, under context e, we measure this semantic scale by the spherical distance between the corrected base prompt and its semantic counterpart. Letting $\hat{u}_{\phi,g} := \mathrm{norm}(Au_{\phi,g} + b)$ and $\hat{u}'_{a,\phi,g} := \mathrm{norm}(Au'_{a,\phi,g} + b)$, and writing $d_{\mathbb{S}}(p, q) := \arccos\langle p, q\rangle$ for the geodesic distance on the shared space $\mathbb{S}^{d-1}$, we define $\ell_a^{(e)} := \mathbb{E}_{(\phi,g)\in\Xi_a}[d_{\mathbb{S}}(\hat{u}_{\phi,g}, \hat{u}'_{a,\phi,g})]$ and $\bar{\tau}_a^{(e)}(x) := \tau_a^{(e)}(x)/\ell_a^{(e)}$. The normalized margin rescales the boundary margin by this context-specific factor, facilitating comparison across contexts with different semantic scales.

Aggregating the normalized margins gives the context-level score

$$S_a(e) := \mathbb{E}_{(x,y)\sim\mathcal{D}_e}\left[\bar{\tau}_a^{(e)}(x) \mid \hat{y} = y\right]. \tag{9}$$

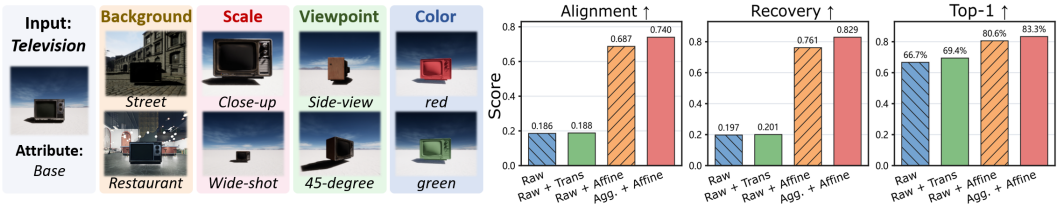
Here, $S_a(e)$ is the expected normalized semantic robustness margin of a under context e, computed over correctly classified samples. In practice, Eq. 9 is estimated by the empirical mean over the corresponding evaluation set $\mathcal{D}_e$. This formulation allows semantic robustness to be compared across contexts through boundary margins. Moreover, the audit does not treat the constructed direction as an oracle semantic direction. Appx. A.3 quantifies how bounded residual directional error propagates to the margins, with explicit lower and upper bounds and empirical stability analysis.

5 Experiments

In this section, we first evaluate the language-derived semantics and then audit VLM robustness.

5.1 Controlled Validation of Semantic Directions

Controlled Visual Validation. We first examine whether the language-derived semantic directions correspond to controlled changes in attributes on the visual side. We use PUG: ImageNet (Bordes et al., 2023), which renders the same object assets under controlled attribute variations. We evaluate 20 classes across 12 controlled attributes, such as background, color, scale, and viewpoint, as illustrated in Fig. 5(a). For each attribute, we derive a visual reference direction by aggregating the embedding differences of controlled image pairs and compare it with the corresponding language-derived direction. We use three complementary measures: Alignment measures similarity between the language-derived direction and its visual reference, Recovery compares this alignment with the agreement between visual directions, and Top-1 tests whether the visual reference identifies the corresponding attribute among all audited directions (see Appx. A.2 for detailed metric definitions).



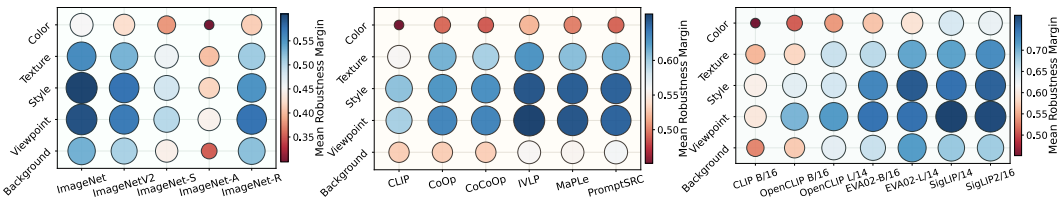
(a) Controlled semantic variations in PUG. (b) Direction construction ablation.

Figure 5: Controlled visual validation on PUG: ImageNet. (a) Representative image pairs of the same *television* asset differing in the target attribute. (b) Alignment, Recovery, and Top-1 under progressively stronger semantic-direction estimation and cross-modal correction.

Visual Alignment. Figure 5(b) traces visual correspondence from a single base prompt through textual aggregation, translation, and affine correction. Aggregation improves correspondence, while translation provides only modest additional gains, consistent with the limited effect of mean-centering in SpLiCE (Bhalla et al., 2024), whereas affine correction substantially improves all three measures. The full estimator reaches 0.74 Alignment, 0.83 Recovery, and 83% Top-1, showing strong alignment with controlled visual semantic variations. These results support using the estimated directions as a reliable semantic basis for auditing. Notably, PUG provides controlled attribute variations rather than an oracle for semantic variation in real images, indicating that the residual discrepancy does not solely reflect estimation error.

5.2 Auditing Across Contexts

Data Shift. With the semantic directions validated, we next audit semantic robustness across data shifts. We compare ImageNet (Deng et al., 2009), ImageNetV2 (Recht et al., 2019), ImageNet-



(a) Auditing under Data Shifts. (b) Auditing under Prompt Learning. (c) Auditing under Model Shifts.

Figure 6: Semantic robustness auditing across (a) data, (b) prompt, and (c) model shifts. Each marker represents one attribute under one evaluation context, with its color and size indicating the mean normalized semantic margin. Larger values indicate greater robustness.

S (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a). Full settings for all auditing experiments are provided in Appx. A.2. Figure 6(a) shows that semantic robustness varies substantially across datasets for the same audited attributes. Across datasets, style and viewpoint are most robust to shifts, while color is consistently the most fragile. ImageNet also yields the largest overall margins, broadly matching the accuracy trend at the dataset level, while the attribute-wise margins expose robustness structure that accuracy does not reveal. By contrast, ImageNet-A and ImageNet-S yield the smallest robustness margins, showing that these shifts place more samples near decision boundaries, consistent with the greater fragility of adversarial and hard examples. This suggests that semantic robustness is strongly shaped by the form of distribution shift rather than preserved uniformly across datasets.

Prompt Shift. We next audit robustness under different prompt learning techniques. We compare the hand-crafted CLIP prompt with CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), IVLP (Khattak et al., 2023a), MaPLe (Khattak et al., 2023a), and PromptSRC (Khattak et al., 2023b) on ImageNet. Figure 6(b) shows that prompt design induces clear changes in semantic robustness. Learned prompts consistently improve over the hand-crafted baseline on texture, style, viewpoint, and background. IVLP achieves the strongest robustness on style and viewpoint, while PromptSRC remains competitive across attributes. Despite these gains, color remains the weakest attribute, which indicates that prompt learning can strengthen semantic robustness overall, but mainly amplifies existing attribute structure rather than eliminating it.

Model Shift. We further audit robustness across VLMs on ImageNet, covering CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), EVA02 (Sun et al., 2023), SigLIP (Zhai et al., 2023), and SigLIP2 (Tschannen et al., 2025). Figure 6(c) shows that semantic robustness is strongly model dependent. Larger or more recent VLMs yield larger margins across most attributes. For example, SigLIP models achieve the strongest overall margins, but their advantage is not uniform, with color still weak in SigLIP2/16 and background less robust in SigLIP/14. This indicates that model scaling improves semantic robustness overall, but preserves attribute-specific vulnerabilities.

5.3 Diagnostic Analysis

Beyond overall auditing across contexts, the proposed framework also supports robustness diagnosis by revealing how semantic robustness is distributed across classes and how boundary-crossing alternatives concentrate into structured confusion patterns.

Class-level Diagnosis. A main advantage of the proposed framework is that it supports analysis beyond attribute-level averages. We next examine semantic robustness at the class level. Figure 7 selects the top-10 classes with the largest mean semantic margins in one setting and tracks their margins across the other settings. Figure 8 gives the complementary view by selecting the top-10 most frequent confusion targets in the first setting and comparing their confusion shares across the remaining settings. Here, confusion share is the fraction of first boundary-crossing events targeting

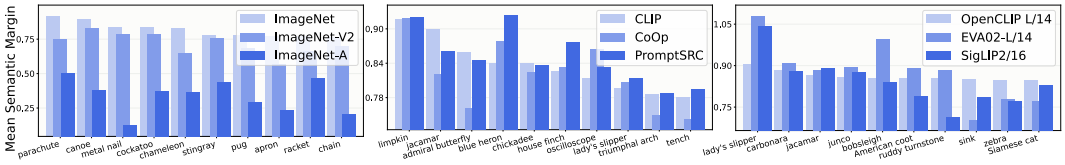
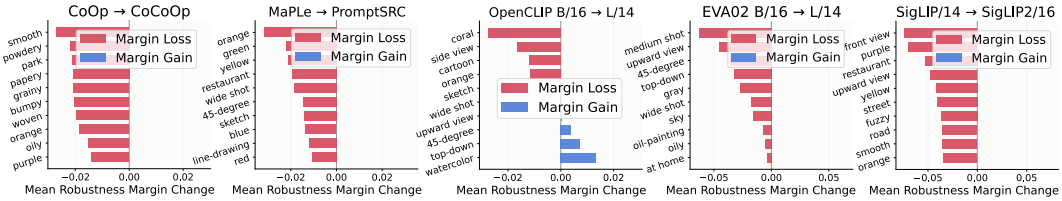


Figure 7: Top-10 robust classes under different auditing contexts. For each comparison, we select the classes with the largest mean semantic margins in the first setting. Bars show their margins across settings, highlighting context-dependent robustness changes.

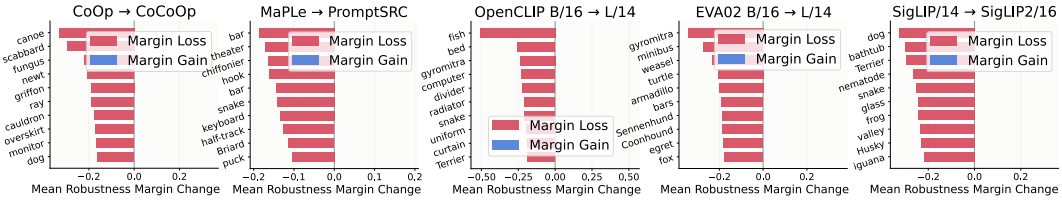


Figure 8: Top-10 confusion targets under different auditing contexts. For each comparison, we select the classes with the largest confusion shares in the first setting. Bars show their shares across settings, highlighting context-dependent confusion changes.



(a) Comparison of Prompt Learning. (b) Comparison of VLMs.

Figure 9: Scaling is not uniform across attributes. For each matched pair, we report the 10 attributes with the lowest signed change in mean normalized margin. Bars show the robustness margin change from the first to second setting, with positive and negative values indicating margin gains and losses.



(a) Comparison of Prompt Learning. (b) Comparison of VLMs.

Figure 10: Scaling is not uniform across classes. For each matched pair, we report the 10 classes with the lowest signed change in mean normalized margin. Bars show the robustness margin change from the first to second setting, with positive and negative values indicating margin gains and losses.

each competing class. These results reveal substantial class-level heterogeneity hidden by global averages. Classes that are highly robust or frequently confused in one setting can change markedly across auditing contexts. These shifts show that the class-level audit can serve as a practical diagnostic, revealing context-dependent class-specific robustness patterns hidden behind aggregate scores.

Non-uniform Robustness Scaling. We next examine whether robustness scaling is uniform under stronger prompt learners or larger model backbones. We consider matched pairs within each family, covering OpenCLIP B/16 to OpenCLIP L/14, EVA02-B/16 to EVA02-L/14, SigLIP/14 to SigLIP2/16, CoOp to CoCoOp, and MaPLe to PromptSRC. For each matched pair, we compute the change in mean normalized semantic margin from the first setting to the second. Figure 9 reports the 10 attributes with the lowest margin changes, and Figure 10 reports the corresponding class-level results. Positive changes indicate margin gains, while negative changes indicate margin losses. Although scaling often improves the overall semantic margin, the gains are not uniformly distributed. At the attribute level, the worst-case behavior differs substantially across scaling families, where OpenCLIP retains margin gains even among some of its least-improved attributes, whereas SigLIP2 exhibits margin losses on several attributes, indicating that stronger overall robustness can still be accompanied by attribute-specific sacrifices. The class-level results show stronger non-uniformity. Across both prompt and model scaling pairs, the worst-affected classes are almost uniformly associated with margin losses, showing that robustness scaling can improve overall robustness while degrading a subset of vulnerable categories. These worst-case changes reveal non-uniform robustness scaling that is invisible from aggregate comparisons alone. Further applications in Appx. A.5 show that class and attribute robustness are broadly consistent with task priors, while model scaling affects robustness differently across domains, making model selection depend on target requirements.

6 Conclusion

This work investigates whether language can provide a reliable representation of semantic variation for auditing VLM robustness. By aggregating semantic directions across textual contexts and calibrating them across modalities, we construct language-derived representations that align well with corresponding visual semantic variations. On this semantic basis, we derive closed-form boundary margins for comparing semantic robustness across data, prompt, and model contexts. Our audits reveal structured robustness variation, class-level disparities, and non-uniform scaling that are not captured by overall performance. Domain-specific analyses further show that the preferred model depends on the target domain. Overall, semantic robustness auditing provides a principled and interpretable perspective for understanding the fine-grained semantic robustness of VLMs.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of the manuscript. GPT-5.4 was additionally used in the explicitly described experiments to obtain fine-grained class domains in Sec. 4.1 and domain priors in Appx. A.5. These uses do not contribute to the development of the proposed method or theoretical analysis. All AI-assisted content and outputs were reviewed by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36:45020–45054, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15702–15712, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Reza Esfandiarpoor, Cristina Menghini, and Stephen Bach. If clip could talk: Understanding vision-language model representations through their preferred concept descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9797–9819, 2024.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor I Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, et al. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Siwon Kim, Jinoh Oh, Sungjin Lee, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Grounding counterfactual explanation of image classifiers to textual concept space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10942–10950, 2023.

Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Hemm: Holistic evaluation of multimodal foundation models. *Advances in Neural Information Processing Systems*, 37:42899–42940, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *arXiv preprint arXiv:2504.10514*, 2025.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, pp. 31716–31731. PMLR, 2023.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *International Joint Conference on Neural Networks*, pp. 1453–1460. IEEE, 2011.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (3):1877–1893, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36:19769–19782, 2023.

A Appendix

A.1 Proof

In this part, we provide the proof of Proposition 4.1.

Proof of Proposition 4.1. Let

$$e_a \coloneqq \frac{1}{n_\phi n_g} \sum_{\phi \in \Phi_a} \sum_{g \in \mathcal{G}_a} d_{a,\phi,g} - m_a.$$

By Eq. 4, the deviation term can be written as

$$e_a = \frac{1}{n_\phi} \sum_{\phi \in \Phi_a} b^{\text{tmp}}_{a,\phi} + \frac{1}{n_g} \sum_{g \in \mathcal{G}_a} b^{\text{cls}}_{a,g} + \frac{1}{n_\phi n_g} \sum_{\phi \in \Phi_a} \sum_{g \in \mathcal{G}_a} \varepsilon_{a,\phi,g}.$$

The centeredness and uncorrelatedness assumptions imply that all cross terms vanish in expectation as

$$\mathbb{E}\|e_a\|_2^2 = \frac{1}{n_\phi^2} \sum_{\phi \in \Phi_a} \mathbb{E}\|b^{\text{tmp}}_{a,\phi}\|_2^2 + \frac{1}{n_g^2} \sum_{g \in \mathcal{G}_a} \mathbb{E}\|b^{\text{cls}}_{a,g}\|_2^2 + \frac{1}{n_\phi^2 n_g^2} \sum_{\phi \in \Phi_a} \sum_{g \in \mathcal{G}_a} \mathbb{E}\|\varepsilon_{a,\phi,g}\|_2^2,$$

and therefore

$$\mathbb{E}\|e_a\|_2^2 \le \frac{\sigma^2_{\text{tmp}}}{n_\phi} + \frac{\sigma^2_{\text{cls}}}{n_g} + \frac{\sigma^2_\varepsilon}{n_\phi n_g}.$$

By definition, the aggregated direction satisfies

$$d_a = \text{norm}(m_a + e_a).$$

Since both d_a and m_a have unit norm,

$$1 - \langle d_a, m_a \rangle = \frac{1}{2}\|d_a - m_a\|_2^2.$$

Moreover, the distance between d_a and m_a can be bounded as

$$\|d_a - m_a\|_2 \le \|d_a - (m_a + e_a)\|_2 + \|e_a\|_2 = \big|1 - \|m_a + e_a\|_2\big| + \|e_a\|_2 \le 2\|e_a\|_2,$$

where the last step uses $\big|\|m_a + e_a\|_2 - \|m_a\|_2\big| \le \|e_a\|_2$ and $\|m_a\|_2 = 1$.

Combining the two displays above gives

$$1 - \langle d_a, m_a \rangle \le 2\|e_a\|_2^2.$$

Taking expectation and combining with the bound on $\mathbb{E}\|e_a\|_2^2$ proves Eq. 5. □

A.2 Experimental Setup

We use OpenAI CLIP ViT-B/16 as the default backbone unless otherwise specified. All experiments were conducted on a single NVIDIA H100 GPU with 80 GB HBM3 memory, dual Intel Xeon Platinum 8462Y+ CPUs with 64 cores in total. The implementation is based on PyTorch and the official implementations of the evaluated prompt-learning methods and VLM backbones.

For ImageNet-based prompt learning experiments, we construct a fixed class-balanced half split of the ImageNet validation set. One half is used for prompt adaptation, and the other half is used for evaluation. All prompt-learning baselines use the same CLIP ViT-B/16 backbone and are trained on the adaptation split using the official default configuration of each method, without additional hyperparameter tuning for our audit. This includes the official optimizer, learning-rate schedule, context length, prompt depth, and training epochs for each method. Images are resized to 224×224 using bicubic interpolation and normalized with the standard CLIP preprocessing unless otherwise specified. SigLIP2 is evaluated at its native 384×384 input resolution.

Table 1: Semantic attribute values used in the ImageNet-family audits. Each value is paired with ImageNet-1K class contexts to construct text-contrast directions within its attribute family.

Attribute	#Values	Attribute values
Color	12	red, orange, coral, yellow, green, blue, purple, pink, brown, gray, white, black
Texture	12	smooth, rough, grainy, bumpy, fuzzy, woven, jagged, cracked, slimy, oily, papery, powdery
Style	6	sketch, line-drawing, cartoon, watercolor, oil-painting, 3d-render
Viewpoint	8	top-down, 45-degree, front view, side view, upward view, close-up, medium shot, wide shot
Background	10	at home, in a kitchen, in a restaurant, on a table, on a street, on a road, in a parking lot, in a garden, in a park, in the sky

Table 2: Default prompt templates for semantic direction construction. Here $\langle c \rangle$ denotes the class name and $\langle a \rangle$ denotes the attribute value.

Attribute	Placement	Prompt templates
Color	Pre-class	a photo of a $\langle a \rangle$ $\langle c \rangle$; an image of a $\langle a \rangle$ $\langle c \rangle$; a photo of the $\langle a \rangle$ $\langle c \rangle$; this is a $\langle a \rangle$ $\langle c \rangle$; a close-up photo of a $\langle a \rangle$ $\langle c \rangle$
Texture	Post-class	a photo of a $\langle c \rangle$ $\langle a \rangle$; an image of a $\langle c \rangle$ $\langle a \rangle$; a close-up photo of a $\langle c \rangle$ $\langle a \rangle$; a picture of a $\langle c \rangle$ $\langle a \rangle$
Style	Pre-class / post-class	a $\langle a \rangle$ photo of a $\langle c \rangle$; a $\langle a \rangle$ style image of a $\langle c \rangle$; a picture of a $\langle c \rangle$ in $\langle a \rangle$ style
Viewpoint	Pre-class / post-class	a $\langle a \rangle$ photo of a $\langle c \rangle$; a $\langle a \rangle$ image of a $\langle c \rangle$; a photo of a $\langle c \rangle$ from the $\langle a \rangle$ view
Background	Post-class	a photo of a $\langle c \rangle$ $\langle a \rangle$; an image of a $\langle c \rangle$ $\langle a \rangle$; a picture of a $\langle c \rangle$ $\langle a \rangle$; a close-up photo of a $\langle c \rangle$ $\langle a \rangle$

Tables 1 and 2 report the semantic attribute values and default prompt templates used to construct directions in the ImageNet-family audits. Each attribute value is paired with ImageNet-1K class contexts, and the resulting text-contrast directions are aggregated over templates and categories following Sec. 4.1. Table 2 also records the attribute placement rule, where color is inserted before the class phrase, texture and background after the class phrase, while style and viewpoint use both pre-class and post-class realizations.

Cross-modal correction and auditing. For each VLM, we fit a single affine correction on the adaptation half of the fixed ImageNet split described above, using each image and its corresponding label text as a paired sample. Let $z_i = f_{\text{img}}(x_i)$ and $u_i = f_{\text{text}}(t_{y_i})$. We estimate

$$(A, b) = \arg\min_{A,b} \sum_i \|z_i - (Au_i + b)\|_2^2. \tag{10}$$

The fitted mapping is fixed for all audits using the same VLM. Thus, the same CLIP correction is used across data and prompt settings, while each VLM in the model comparison is fitted separately on the same ImageNet split. We set $\delta_{\max} = X$ for all auditing experiments. Sample-level margins follow Eq. 8, with context-specific normalization defined in Sec. 4.3. Context-level scores are then obtained by averaging over correctly classified evaluation samples as in Eq. 9.

Controlled visual validation. We use PUG: ImageNet (Bordes et al., 2023), which renders the same object assets under controlled attribute variations. We evaluate 20 classes across 12 controlled attributes spanning background, color, scale, and viewpoint. Let $\mathcal{A}$ denote the evaluated attributes and $\mathcal{P}_a$ the matched image pairs for attribute a. We construct the visual reference direction as

$$v_a = \text{norm}\Big(\frac{1}{|\mathcal{P}_a|} \sum_{(x,x') \in \mathcal{P}_a} [f_{\text{img}}(x') - f_{\text{img}}(x)]\Big). \tag{11}$$

Let $\tilde{d}_a$ denote the corresponding language-derived direction. Alignment is

$$\text{Alignment} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \langle v_a, \tilde{d}_a \rangle. \tag{12}$$

For Recovery, we randomly divide the object assets into two disjoint subsets and independently estimate visual directions $v_{a,r}^{(1)}$ and $v_{a,r}^{(2)}$ over 100 random splits. We define the visual agreement as

$$q_a = \frac{1}{100}\sum_{r=1}^{100}\langle v_{a,r}^{(1)}, v_{a,r}^{(2)}\rangle, \quad \text{Recovery} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{\langle v_a, \tilde{d}_a\rangle}{q_a}. \tag{13}$$

Top-1 measures whether the visual reference retrieves its corresponding semantic direction:

$$\text{Top-1} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\mathbf{1}[a = \arg\max_{b\in\mathcal{A}}\langle v_a, \tilde{d}_b\rangle]. \tag{14}$$

A.3 Reliability of Auditing under Directional Uncertainty

In Section 4.3, we derive a closed-form margin for semantic robustness auditing using the constructed semantic direction. The reliability of this audit depends on whether the direction remains stable under residual contextual inconsistency and cross-modal misalignment. We therefore analyze the effect of directional uncertainty from both theoretical and empirical perspectives. First, we derive a finite residual bound showing that the target semantic margin remains confined to an explicit interval. Second, we empirically examine whether the audit conclusions remain stable as directional uncertainty increases.

A.3.1 Margin Bound under Directional Residuals

We first analyze how bounded directional residuals affect the computed semantic margin. Let d_a° denote the oracle visual semantic direction for attribute a. We assume that the corrected direction $\tilde{d}_a$ satisfies

$$\|d_a^\circ - \tilde{d}_a\|_2 \le \omega_a, \tag{15}$$

where $\omega_a \ge 0$ denotes the remaining residual budget.

For a sample x, let $\hat{y} = f(z)$ be the predicted class. Recall from Eq. 7 that, under the semantic auditing model $\gamma_a(\delta; z)$, the pairwise margin between $\hat{y}$ and a competing class c is given by

$$m_{\hat{y},c}(\delta; x) = \frac{\alpha_{\hat{y},c} + \delta\beta_{\hat{y},c}}{\|z + \delta\tilde{d}_a\|_2}.$$

Since the denominator is strictly positive, the first boundary crossing is governed by the numerator $\alpha_{\hat{y},c} + \delta\beta_{\hat{y},c}$. We therefore write

$$\Delta_{\hat{y},c} \coloneqq u_{\hat{y}} - u_c, \quad \alpha_{\hat{y},c} \coloneqq \langle z, \Delta_{\hat{y},c}\rangle, \quad \beta_{\hat{y},c} \coloneqq \langle \tilde{d}_a, \Delta_{\hat{y},c}\rangle.$$

Here, $\Delta_{\hat{y},c}$ is the normal direction of the pairwise decision boundary between $\hat{y}$ and c, $\alpha_{\hat{y},c}$ is the original pairwise decision margin at z, and $\beta_{\hat{y},c}$ measures how fast this margin changes along the corrected semantic direction $\tilde{d}_a$.

Under the visual direction d_a°, the same pairwise boundary has the slope

$$\beta_{\hat{y},c}^\circ \coloneqq \langle d_a^\circ, \Delta_{\hat{y},c}\rangle.$$

Thus, directional residual does not change the initial margin $\alpha_{\hat{y},c}$, but only perturbs the slope from $\beta_{\hat{y},c}$ to $\beta_{\hat{y},c}^\circ$. By Eq. 15 and Cauchy–Schwarz,

$$\left|\beta_{\hat{y},c}^\circ - \beta_{\hat{y},c}\right| = \left|\langle d_a^\circ - \tilde{d}_a, \Delta_{\hat{y},c}\rangle\right| \le \omega_a\|\Delta_{\hat{y},c}\|_2. \tag{16}$$

We denote this class-dependent slope uncertainty by

$$\kappa_{\hat{y},c} \coloneqq \omega_a\|\Delta_{\hat{y},c}\|_2.$$

Since the class embeddings are normalized, $\|\Delta_{\hat{y},c}\|_2 \le 2$, and hence $\kappa_{\hat{y},c} \le 2\omega_a$.

The bound in Eq. 16 implies that, for each competing class c, the slope induced by the oracle visual direction satisfies

$$\beta_{\hat{y},c}^\circ \in [\beta_{\hat{y},c} - \kappa_{\hat{y},c},\ \beta_{\hat{y},c} + \kappa_{\hat{y},c}].$$

A boundary crossing along the positive semantic extent occurs only when the corresponding slope is negative. Therefore, the earliest admissible crossing for class c is obtained by taking the most negative slope, while a crossing is guaranteed for class c only if even the least negative admissible slope remains below zero. Over the finite auditing range $[0, \delta_{\max}]$, this leads to the following residual-aware lower and upper margins:

$$\tau_a^-(x) := \min\left\{\delta_{\max}, \min_{c\neq\hat{y}:\, \beta_{\hat{y},c}-\kappa_{\hat{y},c}<0} \frac{\alpha_{\hat{y},c}}{\kappa_{\hat{y},c}-\beta_{\hat{y},c}}\right\}, \tag{17}$$

and

$$\tau_a^+(x) := \min\left\{\delta_{\max}, \min_{c\neq\hat{y}:\, \beta_{\hat{y},c}+\kappa_{\hat{y},c}<0} \frac{\alpha_{\hat{y},c}}{-\beta_{\hat{y},c}-\kappa_{\hat{y},c}}\right\}. \tag{18}$$

If no competing class satisfies the corresponding slope condition, the margin is set to $\delta_{\max}$, indicating stability up to the audited range endpoint.

These definitions convert the directional residual bound into an interval-valued robustness margin. The following proposition formalizes that any margin induced by an admissible oracle direction must lie within this interval.

Proposition A.1. *Suppose Eq. 15 holds. Let $\tau_a^\circ(x)$ denote the semantic boundary margin computed along the oracle visual direction d_a°. Then the oracle semantic margin satisfies*

$$\tau_a^-(x) \le \tau_a^\circ(x) \le \tau_a^+(x). \tag{19}$$

Proof. For each competing class $c \neq \hat{y}$, the pairwise boundary under the oracle visual direction is reached when

$$\alpha_{\hat{y},c} + \delta\beta_{\hat{y},c}^\circ = 0.$$

Since $\hat{y} = f(z)$, we have $\alpha_{\hat{y},c} \ge 0$ for all $c \neq \hat{y}$. Thus, a boundary crossing along the positive semantic extent can occur only when $\beta_{\hat{y},c}^\circ < 0$, in which case the crossing point is

$$\tau_{\hat{y},c}^\circ = -\alpha_{\hat{y},c}/\beta_{\hat{y},c}^\circ.$$

If $\beta_{\hat{y},c}^\circ \ge 0$, no positive-direction crossing occurs within the finite auditing range.

By Eq. 16, the oracle slope satisfies

$$\beta_{\hat{y},c} - \kappa_{\hat{y},c} \le \beta_{\hat{y},c}^\circ \le \beta_{\hat{y},c} + \kappa_{\hat{y},c}.$$

For a fixed class c, the earliest admissible crossing is obtained by taking the most negative admissible slope, $\beta_{\hat{y},c} - \kappa_{\hat{y},c}$. Therefore, whenever $\beta_{\hat{y},c} - \kappa_{\hat{y},c} < 0$, any admissible oracle crossing for this class is no earlier than

$$\frac{\alpha_{\hat{y},c}}{\kappa_{\hat{y},c}-\beta_{\hat{y},c}}.$$

Minimizing over competing classes and capping by $\delta_{\max}$ yields $\tau_a^-(x) \le \tau_a^\circ(x)$.

For the upper bound, if $\beta_{\hat{y},c} + \kappa_{\hat{y},c} < 0$, then the slope remains negative for all admissible oracle directions. Hence class c necessarily induces a crossing within the positive direction, unless it lies beyond the finite auditing range, and its latest admissible crossing is obtained at the least negative admissible slope, $\beta_{\hat{y},c} + \kappa_{\hat{y},c}$. Thus, for such a class, the oracle crossing point is bounded as

$$\tau_{\hat{y},c}^\circ \le \frac{\alpha_{\hat{y},c}}{-\beta_{\hat{y},c}-\kappa_{\hat{y},c}}.$$

Taking the minimum over all classes satisfying $\beta_{\hat{y},c} + \kappa_{\hat{y},c} < 0$ and capping by $\delta_{\max}$ gives $\tau_a^\circ(x) \le \tau_a^+(x)$. Combining the two inequalities establishes Eq. 19 and proves the proposition. □

The interval in Eq. 19 quantifies the stability of the semantic margin under residual directional uncertainty. Its width is controlled by the class-dependent slope uncertainty $\kappa_{\hat{y},c}$, which scales with the residual budget ω_a. When ω_a is small, the admissible slope interval in Eq. 16 is narrow, and the residual-aware margin remains close to the margin computed from $\tilde{d}_a$. In contrast, when $\beta_{\hat{y},c}$ is close to zero for a competing class, even a small residual may change whether that class moves toward or away from the decision boundary. The resulting wider interval therefore identifies cases where the audit margin is more sensitive to directional uncertainty.

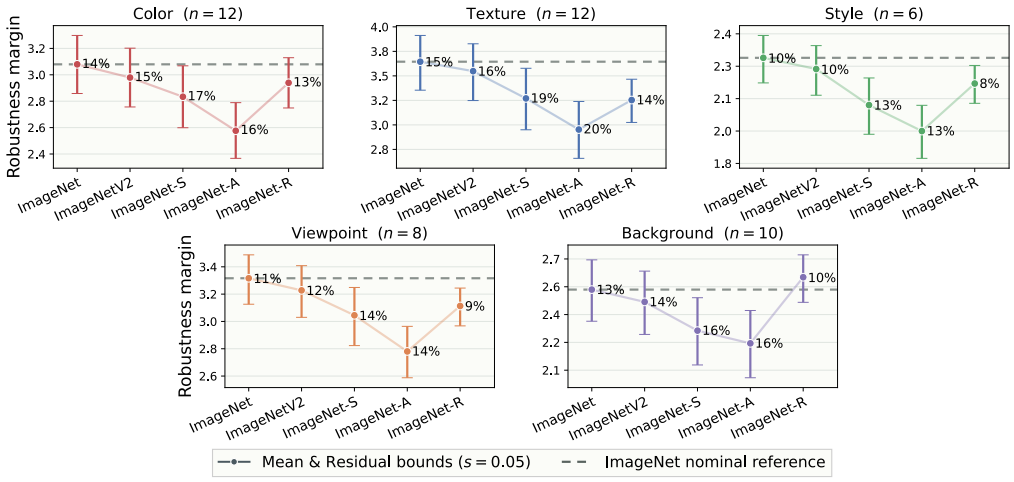
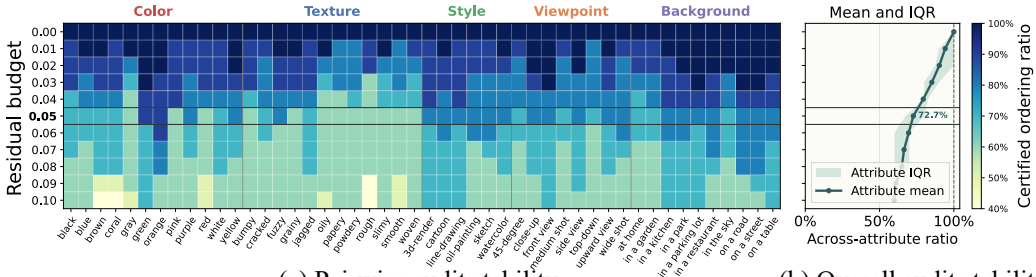


Figure 11: Residual-aware robustness intervals under directional uncertainty. Points show the mean normalized semantic margins across ImageNet-family datasets at $s = 0.05$, with error bars indicating the residual-aware lower and upper scores. Percentages report interval width relative to the nominal score, and dashed lines mark the ImageNet nominal reference.



(a) Pairwise audit stability. (b) Overall audit stability.

Figure 12: Stability of audit conclusions under directional uncertainty. (a) Certified ordering ratio for each attribute as the residual budget increases. An original ordering is certified when the lower bound of the higher-ranked context remains above the upper bound of the lower-ranked context. (b) Average certified ordering ratio across attributes, with the shaded region showing the middle 50% of attributes. Higher values indicate that more original cross-context orderings remain guaranteed.

The same bound applies to the normalized margin used for cross-context auditing. Since $\ell_a^{(e)} > 0$, dividing Eq. 19 by the context-dependent semantic scale gives

$$\bar{\tau}_a^-(x) \le \bar{\tau}_a^\circ(x) \le \bar{\tau}_a^+(x), \quad \bar{\tau}_a^-(x) := \frac{\tau_a^-(x)}{\ell_a^{(e)}}, \quad \bar{\tau}_a^+(x) := \frac{\tau_a^+(x)}{\ell_a^{(e)}}. \tag{20}$$

This provides a theoretical reliability guarantee for the auditing margin under bounded residual directional error, showing that the effect of residual uncertainty remains explicitly controlled on the same normalized semantic scale. Aggregating τ_a^- and τ_a^+ as in Eq. 9 gives the corresponding context-level lower and upper scores $S_a^-(e)$ and $S_a^+(e)$.

A.3.2 Empirical Sensitivity under Directional Uncertainty

We next empirically examine whether the audit conclusions remain stable under directional uncertainty. We focus on the ImageNet data-shift setting. Since the oracle direction is unknown, we use the observed dispersion of corrected contextual directions as a reference residual scale. Let $\tilde{d}_{a,\xi}$ denote the corrected direction for textual context ξ before aggregation and $\tilde{d}_a$ the final aggregated direction. We define

$$\hat{\omega}_a = \max_{\xi \in \Xi_a} \|\tilde{d}_{a,\xi} - \tilde{d}_a\|_2. \tag{21}$$

We then vary the residual budget as

$$\omega_a(s) = s\hat{\omega}_a, \tag{22}$$

where s controls the uncertainty level. The guarantees below are therefore conditional on the assumed budget $\omega_a(s)$. For each value of s, we compute the lower and upper margins in Eqs. 17-18, normalize them by $\ell_a^{(e)}$, and aggregate them into context-level score intervals.

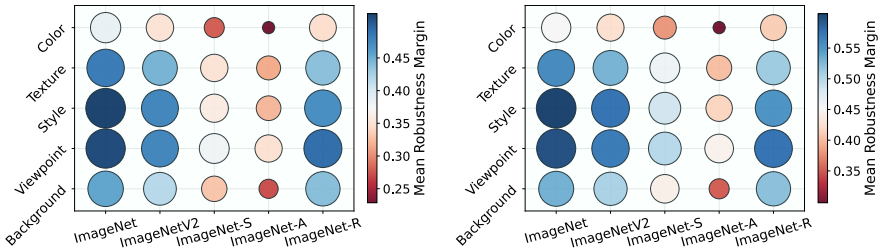
Figure 11 directly compares the nominal audit scores with their residual-aware intervals at $s = 0.05$. Across attribute families, the intervals remain relatively narrow, with relative widths ranging from about 8% to 20%, while the broad cross-dataset robustness structure is preserved. In particular, datasets with substantially different nominal margins remain clearly separated in most cases. This shows that the residual-aware bounds remain informative rather than becoming vacuous under moderate directional uncertainty.

We further examine whether the residual-aware intervals preserve the relative audit conclusions across contexts. An original pairwise ordering is regarded as certified when the lower score of the higher-ranked context remains above the upper score of the lower-ranked context. The certified ordering ratio is the fraction of original context pairs satisfying this condition. Figure 12(a) reports this ratio for each attribute as the residual budget increases, revealing gradual degradation with clear attribute-dependent variation. Figure 12(b) summarizes the ratio across attributes. At $s = 0.05$, 72.7% of the original pairwise orderings remain certified, showing that a substantial fraction of the original cross-context conclusions remains guaranteed under moderate directional uncertainty.

Overall, these results substantiate the theoretical bound in Proposition A.1. The audit margins are affected by residual directional uncertainty, but the effect remains controlled, and the main context-level audit conclusions are largely preserved. This indicates that the proposed semantic robustness audit remains informative under bounded directional uncertainty.

A.4 Auditing Beyond Correctly Classified Samples

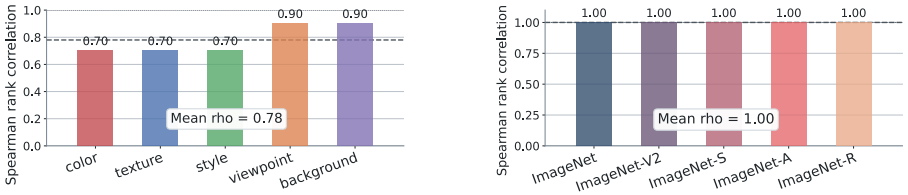
In the main experiments, we report semantic robustness statistics on correctly classified samples. This choice focuses the evaluation on prediction reliability, since the measured margin then characterizes how stable a correct prediction remains under controlled semantic variation. However, this protocol also restricts the use of the audit to labeled benchmarks where correctness can be determined. In many deployment settings, ground-truth labels may be unavailable, while auditing the robustness structure of model predictions remains important.



(a) Auditing over all samples. (b) Auditing over correctly classified samples.

Figure 13: Comparison between all-image and correct-only semantic robustness audits on ImageNet-family datasets. (a) Mean normalized semantic margins computed over all images. (b) Mean normalized semantic margins computed over correctly classified samples. The two maps exhibit similar attribute-dataset structures, showing that the all-image audit preserves the dominant robustness patterns observed under the correct-only protocol.

A key property of our formulation is that the sample-level semantic margin does not require the ground-truth label. The pairwise margin in Eq. 7 and the resulting boundary margin in Eq. 2 are computed with respect to the model prediction $\hat{y} = f(z)$ and its first decision-boundary crossing to any competing class. Therefore, labels are only needed to restrict the aggregation in Eq. 9 to correctly classified samples, not to compute the semantic margin itself. This allows a label-free all-image variant of the audit. Specifically, we use the same normalized sample-level margin $\bar{\tau}_a^{(e)}(x)$, but aggregate it over all images regardless of whether the prediction is correct. For comparison,



(a) Dataset ranking agreement.

(b) Attribute ranking agreement.

Figure 14: Ranking agreement between all-image and correct-only audits. (a) Spearman rank correlation of dataset rankings for each fixed attribute. (b) Spearman rank correlation of attribute rankings for each fixed dataset. High correlations in both views indicate that extending the audit to all images preserves the relative robustness conclusions across datasets and semantic attributes.

the correct-only protocol used in the main experiments applies the same margin computation after restricting the aggregation to samples with $\hat{y} = y$. Thus, the all-image and correct-only protocols differ only in the set of samples used for aggregation, not in the definition of the semantic margin.

Figure 13 compares the all-image audit with the correct-only audit on ImageNet-family datasets. The two maps show highly similar attribute-dataset structures. Although including all images changes the absolute margins slightly, the dominant robustness pattern remains consistent. In particular, the relative behavior across semantic attributes and data shifts is preserved. This indicates that the robustness patterns observed in the main experiments are not primarily caused by the correct-only aggregation protocol.

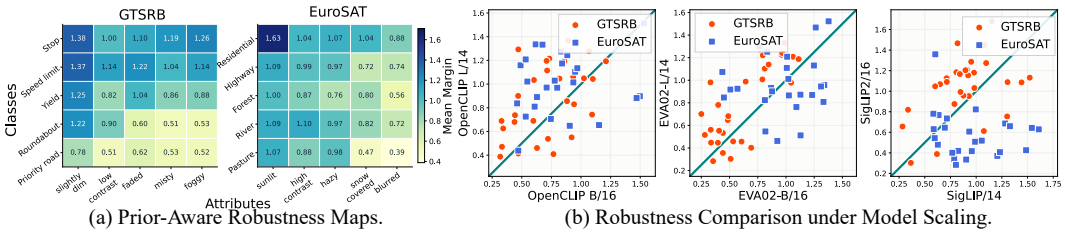
Figure 14 further quantifies this agreement through rank consistency. In Fig. 14(a), we fix each attribute and compare the dataset rankings induced by the all-image and correct-only protocols. The correlations remain high across attributes, showing that the relative robustness of different datasets is largely preserved under auditing. In Fig. 14(b), we fix each dataset and compare the attribute rankings induced by the two protocols. The agreement is nearly perfect across datasets, showing that the semantic robustness ordering over attributes remains stable when all images are included.

Overall, these results show that the proposed audit can be applied in label-free settings while preserving the main comparative conclusions obtained from correctly classified samples. Thus, the all-image audit provides a practical extension for unlabeled data, and the main robustness findings are not artifacts of the correct-only evaluation protocol.

A.5 Application: Domain Attribute Checks and Model Selection

We further demonstrate two practical uses of our auditing framework on GTSRB (Stallkamp et al., 2011) and EuroSAT (Helber et al., 2019): auditing domain-specific robustness under task-relevant visibility attributes and selecting models under domain robustness requirements. For each dataset, we consider five representative classes and five domain-specific attributes related to visibility.

Domain attribute checks. Figure 15(a) uses domain priors derived with GPT-5.4 (Achiam et al., 2023) to order classes and visibility attributes from easy to hard, while the cell values are computed



(a) Prior-Aware Robustness Maps.

(b) Robustness Comparison under Model Scaling.

Figure 15: Case studies on GTSRB and EuroSAT. (a) Mean semantic margins for representative classes and attributes. The class and attribute orders follow a prior robust-to-fragile arrangement, revealing whether the audited patterns align with domain knowledge. (b) Class-attribute robustness under model scaling for three matched model pairs. Each point denotes one class-attribute pair, with the horizontal axis showing the base model and the vertical axis showing the larger model.

solely from our semantic margins. The resulting heatmaps are broadly consistent with these priors. In GTSRB, the audit assigns larger margins to geometrically distinctive signs and smaller margins to fine-grained classes (e.g., *Priority road*) under harder visibility changes. In EuroSAT, spatially distinctive classes such as *Residential* have larger margins, whereas appearance-dominated classes such as *Pasture* are less robust. Thus, the framework localizes domain-specific fragile patterns beyond overall performance, with the agreement to task priors serving as evidence that the audited patterns are meaningful.

Model selection. The audit can guide model selection by checking whether an upgrade improves the class-attribute robustness profile required by a target domain. Figure 15(b) compares each base model with its scaled counterpart at the class-attribute level, where points above the diagonal indicate improved robustness. The three scaling pairs lead to different selection outcomes, with OpenCLIP showing broader cross-domain improvements, EVA02 being more favorable on GTSRB, and SigLIP2 being preferable on GTSRB but often worse on EuroSAT. Therefore, the preferred model depends on the target domain, and model scale alone is insufficient for robust model selection.

A.6 Limitations and Future Directions

The main limitation of our framework is that the semantic margin is exact for the semantic path specified by the constructed auditing direction. As the direction is estimated from textual prompts and calibrated to the visual embedding space through cross-modal correction, the margin should not be interpreted as an absolute guarantee of semantic correctness in the input space. Its interpretation can still be affected by residual directional uncertainty, cross-modal mismatch, and the representation geometry learned by the underlying VLM. We address this uncertainty explicitly rather than leaving it implicit. In Appx. A.3, we provide residual uncertainty bounds that quantify how bounded directional error propagates to the semantic margin, and empirically show that the main context-level conclusions remain stable under controlled directional perturbations. Therefore, the proposed framework should be viewed as a diagnostic and comparative auditing tool with quantified uncertainty, rather than a standalone certificate of real-world reliability.

The framework also suggests several useful extensions for downstream robustness analysis. Appx. A.4 shows that the audit can be applied beyond correctly classified samples, with high agreement between all-image and correct-only results. Since the sample-level semantic margin is computed with respect to the model prediction and does not require ground-truth labels, it can support label-free target-domain auditing, comparison of candidate models on unlabeled data, and identification of samples close to semantic decision boundaries. Another possible extension is to use the more reliable semantic axes constructed in this work as a semantic backend for robustness analysis beyond VLMs. Prior work has shown that textual concepts from VLM spaces can be connected to standard visual classifiers for concept-based and counterfactual analysis (Kim et al., 2023). Our aggregation and cross-modal correction procedures could strengthen such analyses by reducing prompt-specific artifacts and better calibrating text-defined directions to visual representations. For non-VLM classifiers, the resulting axes could be used as semantic probes or mapped to the classifier's latent space, although the closed-form boundary margin would no longer follow directly because it relies on the similarity-based decision geometry of VLMs. Boundary evaluation in this setting would require sampling or optimization along the transferred semantic direction.

A.7 Broader Impact

This work provides a semantic robustness auditing framework for VLMs, aiming to reveal robustness fragilities that are hidden by standard accuracy metrics. By measuring semantic boundary margins, the audit can support fine-grained robustness comparison, class-level diagnosis, target-domain model selection, and label-free auditing when annotations are unavailable or costly. These capabilities may help practitioners identify domain-relevant failure modes before deployment and avoid relying solely on aggregate accuracy when evaluating VLM reliability. However, the audit should be interpreted as diagnostic evidence rather than a deployment certificate. A small or large semantic margin indicates how close a prediction is to a boundary under the constructed auditing model, but it does not by itself guarantee real-world reliability under all possible semantic variations. The results should therefore be interpreted together with the residual uncertainty analysis provided in this work.