GENERATIVE VISION-LANGUAGE MODELS CAN DISCRIMINATE: UNLOCKING ZERO-SHOT CLASSIFICATION POTENTIAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Vision-Language Models (VLMs), such as CLIP, are naturally suited to zero-shot classification through image–text similarity matching, whereas generative VLMs provide an autoregressive decoding interface that is less direct for closed-set classification. Existing protocols based on open-ended generation or candidate-constrained decoding can be sensitive to label wording, output format, and prompt design. This motivates us to examine whether part of the observed performance gap arises from the inference interface rather than representational capacity alone. In this paper, we propose a zero-shot framework that unlocks the discriminative potential of generative VLMs. By inducing a contrastive-compatible latent space directly from the model's internal representations, we establish a similarity-based matching mechanism that mimics the contrastive formulation, while preserving the rich semantics of the generative paradigm. Furthermore, we identify a phenomenon where the VLM's internal linguistic priors accurately recognize visual concepts but diverge from canonical label taxonomies. To mitigate this, we introduce semantic equivalence rectification, which aligns diverse linguistic manifestations to ground-truth categories. Experiments across multiple benchmarks show consistent improvements over generative prompting baselines. Further analyses identify taxonomy–class alignment as an important, but not exclusive, source of SER's gains.

1 INTRODUCTION

Contrastive Vision-Language Models (VLMs), such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021), are used for zero-shot visual classification because their contrastive objectives explicitly align images and text for similarity matching. Generative VLMs, despite their strong open-ended capabilities, are less straightforward to use for this task: their autoregressive outputs can be sensitive to label wording, output format, and prompt design Zhang et al. (2024), as illustrated in Figure 1. As generative VLMs are deployed as general-purpose multimodal backbones, enabling reliable closed-set decisions from the same frozen model would broaden their practical utility.

A key difficulty is the mismatch between open-vocabulary autoregressive decoding and evaluation against a fixed set of benchmark labels. Unlike contrastive models, generative VLMs do not expose an explicitly trained image–text metric space for classification, so their predictions must be converted into closed-set decisions through an inference protocol. Existing protocols generally follow two paradigms: (a) *Open-world generation*. In this setting, the VLM is prompted with instructions such as "Please describe the class in this image," relying on direct open-ended decoding. However, this approach suffers from a severe vocabulary mismatch, where the model projects a probability distribution over a vast vocabulary (often exceeding hundreds of thousands of tokens), and the candidate classes represent a sparse subset of this space. For example, given the visual features of a "Tiger Lily," the model's probability mass frequently disperses across synonymous descriptions such as "orange spotted lily" or superordinate terms like "flower" rather than converging on the precise ground-truth label required by the benchmark. This can produce an incorrect benchmark prediction even when the generated expression is semantically related to the target category. (b) *Closed-world prompting*. This strategy explicitly enumerates the full candidate taxonomy in the input and asks the model to generate one of the provided labels. While practical for compact label sets,

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.

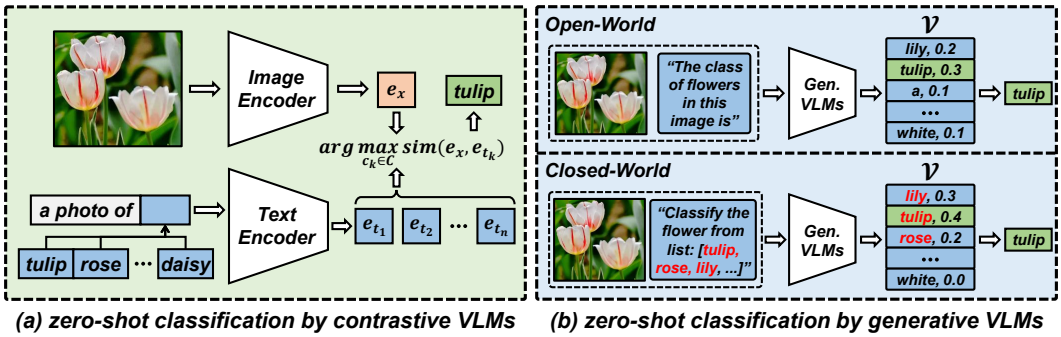

Figure 1: **Comparison of zero-shot classification paradigms. (a) Contrastive VLMs** perform classification by similarity matching between visual and textual embeddings in a shared latent space. **(b) Generative VLMs** perform autoregressive prediction over the vocabulary $\mathcal{V}$. Open-world inference directly generates a textual prediction, whereas closed-world prompting conditions generation on an explicitly enumerated candidate set.

incorporating large taxonomies, such as the 1,000 classes of ImageNet Russakovsky et al. (2015), produces increasingly long prompts and can make predictions sensitive to candidate ordering Haviv et al. (2022). Task-specific fine-tuning may improve classification performance, but falls outside the conventional zero-shot setting Xian et al. (2018). These limitations motivate a representation-level inference interface that does not require the complete candidate list to be repeatedly processed for every input image.

In this work, we revisit the limitations of standard prompting protocols for closed-set classification with generative VLMs Alayrac et al. (2022); Liu et al. (2024a). Rather than attributing the observed errors solely to representational capacity, we examine whether the inference interface itself prevents useful visual information from being expressed as a fixed benchmark label. Because generative architectures do not provide an explicitly contrastive image–text space, we seek a representation-level interface for direct comparison.

As illustrated in Figure 2, we construct a *latent comparable space* from a frozen generative VLM. We extract visual and textual hidden states under matched next-token prediction contexts: the visual prompt prepares the model to predict a class name, while the textual prompt prepares it to predict the same target token from a candidate-class description. Predicting the same token imposes shared constraints through the language modeling head, which encourages the hidden states to share decision-relevant components without guaranteeing complete hidden-state equivalence. We therefore compare their normalized representations by cosine similarity and use the resulting scores for closed-set classification.

We further study *Taxonomic Bias*, where model-preferred expressions may differ from canonical benchmark labels. To reduce such mismatches, Semantic Equivalence Rectification (SER) aggregates scores over aliases, synonyms, and descriptive variants associated with each class. SER uses externally generated text variants and a threshold for adaptive routing, and is therefore reported separately from the primary latent matching method.

Experiments across multiple visual recognition datasets and diverse generative VLM families show that the proposed latent inference formulation consistently improves over a reproduced closed-world prompting baseline. The gains are generally larger when long candidate lists make direct decoding unreliable, but become smaller when recent VLMs already handle compact label sets effectively. These results indicate that latent matching is a useful adaptation for frozen generative VLMs, with its effectiveness depending on the model backbone, candidate taxonomy, semantic-bank quality, and inference calibration.

Our main contributions are summarized as follows:

- We analyze how open-vocabulary decoding, large candidate sets, and benchmark-specific label conventions can limit closed-set classification with generative VLMs.

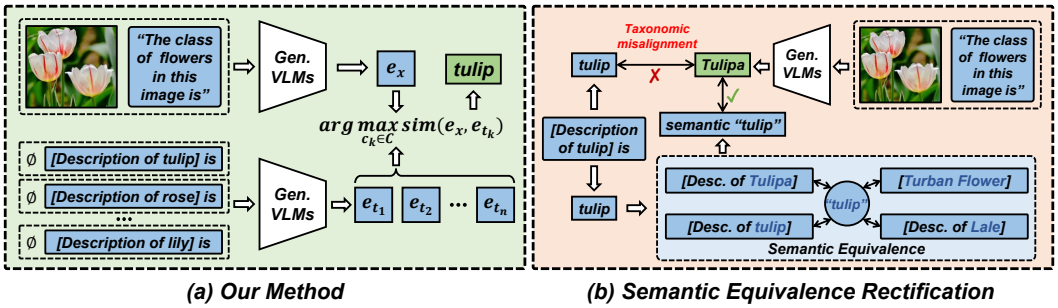

Figure 2: **Overview of our proposed framework. (a) Induced Aligned Space** extracts hidden states from image-prompt inputs and text-only class descriptions ($\emptyset$ denotes no image input), enabling classification via similarity matching. **(b) Semantic Equivalence Rectification** aggregates scores over equivalent terms to map predictions to canonical concepts.

- We propose a frozen-backbone latent matching framework that compares visual and textual hidden states extracted under matched next-token prediction contexts, together with SER for incorporating alternative class expressions.
- We evaluate the framework across multiple datasets and seven generative VLMs, showing consistent improvements over a reproduced closed-world prompting baseline. A controlled audit with wrong-class mappings further indicates that taxonomy–class alignment contributes substantially, but not exclusively, to SER's gains.

2 Related Work

2.1 Contrastive Vision-Language Models and Applications

Contrastive Vision-Language Models (VLMs), such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021), learn aligned image–text representations through multimodal pre-training. Their dual-encoder architecture enables similarity matching between visual embeddings and textual class descriptions, making them well suited to zero-shot classification. Subsequent models, such as BLIP Li et al. (2022b), extend multimodal alignment through unified encoder–decoder architectures. Beyond classification, aligned representations support cross-modal retrieval, open-vocabulary detection Gu et al. (2021); Lüddecke & Ecker (2022), and text-to-image generation guidance Rombach et al. (2022). However, dual-encoder contrastive models do not provide the open-ended generation and dialogue interface offered by generative VLMs Liu et al. (2023; 2024c).

2.2 Limitations of Generative Vision-Language Models

Generative VLMs, such as Flamingo Alayrac et al. (2022) and LLaVA Liu et al. (2023), integrate LLMs with visual encoders to support open-ended instruction following and multimodal dialogue. Despite their broader generative capabilities and often larger parameter counts, these models can underperform contrastive models on standard image classification and fine-grained recognition benchmarks Zhang et al. (2024); Liu et al. (2024c).

Prior work links this gap to open-vocabulary decoding, prompt sensitivity, and difficulties in mapping generated responses to fixed labels Conti et al. (2025); Zhang et al. (2024). Generative models may also hallucinate or overlook fine-grained visual details Guan et al. (2024), indicating that label mismatch is not the only source of classification error. Existing studies have not directly examined whether hidden states from matched prediction contexts provide a useful frozen-model comparison interface, or how taxonomy–class correspondence contributes to semantic expansion. Our work investigates these two questions.

3 METHODOLOGY

3.1 PRELIMINARIES

Zero-shot Classification with Contrastive VLMs. Given an input image $\boldsymbol{x}$ and a set of category labels $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, contrastive Vision-Language Models (e.g., CLIP Radford et al. (2021)) employ visual and textual encoders, denoted as $\mathcal{M}_x$ and $\mathcal{M}_t$ respectively, to project image and text into a shared embedding space $\mathbb{R}^d$. To perform zero-shot classification, each category label c_k is typically instantiated into a descriptive text $\boldsymbol{t}_k$ via a prompt template (e.g., "a photo of a $[c_k]$"). The predicted label $\hat{y}$ is then determined by retrieving the candidate that maximizes the similarity with the image embedding:

$$\begin{aligned} \hat{y} &= \arg\max_{c_k \in \mathcal{C}} \operatorname{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{t_k}\right), \\ \boldsymbol{e}_x &= \mathcal{M}_x(\boldsymbol{x}), \qquad \boldsymbol{e}_{t_k} = \mathcal{M}_t(\boldsymbol{t}_k). \end{aligned} \tag{1}$$

where $\boldsymbol{e}_x$ and $\boldsymbol{e}_{t_k}$ denote the feature representations of the image and the k-th text description, respectively, and $\operatorname{sim}(\cdot, \cdot)$ denotes the similarity function. This paradigm relies on the metric alignment established during pre-training, which pulls semantically paired image-text embeddings closer in the shared space. This alignment between the pre-training objective and downstream matching formulation underpins the success of contrastive VLMs in zero-shot classification.

Zero-shot Classification with Generative VLMs. Unlike the metric-based alignment in contrastive models, generative VLMs, denoted as $\mathcal{M}$, operate in an autoregressive manner. Given an input image $\boldsymbol{x}$ and a prompt $\boldsymbol{t}_p$, the model predicts the next token t_c from its vocabulary $\mathcal{V}$. Formally, let $\boldsymbol{h} \in \mathbb{R}^d$ denote the final hidden-state representation used for next-token prediction, computed as $\boldsymbol{h} = \mathcal{M}(\boldsymbol{x}, \boldsymbol{t}_p)$. The probability distribution over the vocabulary is obtained by projecting $\boldsymbol{h}$ through the language-model output-head matrix $\boldsymbol{W}$:

$$\begin{aligned} P(t_c \mid \boldsymbol{x}, \boldsymbol{t}_p) &= \frac{\exp(\boldsymbol{w}_{t_c}^\top \boldsymbol{h})}{\sum_{t_j \in \mathcal{V}} \exp(\boldsymbol{w}_{t_j}^\top \boldsymbol{h})}, \\ \hat{t} &= \arg\max_{t_c \in \mathcal{V}} P(t_c \mid \boldsymbol{x}, \boldsymbol{t}_p). \end{aligned} \tag{2}$$

where $\boldsymbol{w}_{t_c}$ denotes the output weight vector associated with token t_c. In practice, existing methods adapt this paradigm for zero-shot classification through the following two implementations.

(i) *Open-World Generation.* The prompt $\boldsymbol{t}_p$ is a generic instruction (e.g., "The class of the main object in the image is"). Since next-token prediction is defined over the model vocabulary $\mathcal{V}$ rather than the task-specific candidate set $\mathcal{C}$, the generated output $\hat{y}$ often drifts to synonyms or descriptive terms that do not match the ground-truth label (i.e., $\hat{y} \notin \mathcal{C}$), resulting in failures.

(ii) *Closed-World Generation.* To constrain the output space, the prompt explicitly enumerates the candidates (e.g., "Choose from: $c_1, \ldots, c_K$, the class of the main object in the image is"). This method faces two practical limitations. First, incorporating large-scale taxonomies, such as the 1,000 classes of ImageNet Russakovsky et al. (2015), into a single prompt can strain the available context. Second, predictions can also be sensitive to the permutation order of candidate labels due to the causal attention mechanism Haviv et al. (2022). These limitations can make closed-world generation less reliable as the candidate set grows.

3.2 OUR METHOD

Induced Aligned Space. To obtain a metric-based decision interface from a frozen generative VLM, we use the language modeling head as a shared reference. When different contexts assign high probability to the same target token, their hidden states satisfy related logit-margin constraints in the subspace visible to the output projection. This does not imply that the complete hidden states must converge; instead, it motivates testing whether their normalized representations are sufficiently comparable for classification. The appendix provides the analysis and controlled evidence.

Leveraging this observation, we extract two representations under matched prediction contexts:

(i) *Visual Anchor Induction.* Given an image $\boldsymbol{x}$, a prompt $\boldsymbol{t}_x$ (e.g., "The object in the image is") produces the hidden state $\boldsymbol{e}_x$ immediately before next-token prediction Brown et al. (2020). This representation serves as the visual anchor for subsequent similarity matching.

(ii) *Textual Anchor Induction.* For class c_k, a text-only prompt $\boldsymbol{t}_{c,k}$ (e.g., "[Description of c_k] is a") produces the hidden state $\boldsymbol{e}_{c_k}$ at the corresponding position. Further representation-extraction details are reported in Appendix.

Because representations are extracted under matched target-token prediction contexts, we evaluate their comparability using normalized cosine similarity and cast classification, similar to Eq. (1), as:

$$\begin{aligned} \hat{y} &= \arg\max_{c_k \in \mathcal{C}} \operatorname{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{c_k}\right), \\ \boldsymbol{e}_x &= \mathcal{M}(\boldsymbol{x}, \boldsymbol{t}_x), \qquad \boldsymbol{e}_{c_k} = \mathcal{M}(\emptyset, \boldsymbol{t}_{c,k}). \end{aligned} \tag{3}$$

Semantic Equivalence Rectification. The primary textual anchor is constructed from a canonical benchmark label and may not cover alternative expressions that the model associates with the same visual concept, especially for fine-grained categories. For example, a visual representation associated with the scientific name "Helianthus annuus" may not be sufficiently close to an anchor constructed only from the benchmark label "sunflower." We refer to errors attributable to this mismatch between model-preferred expressions and benchmark labels as *Taxonomic Bias*.

To provide broader textual support, for each category $c_k \in \mathcal{C}$, we construct an augmented semantic bank $\mathcal{S}_k = \{s_{k,1}, \ldots, s_{k,M}\}$ containing aliases, scientific names, common synonyms, and visual descriptions generated from the class name by an external LLM. Each variant $s \in \mathcal{S}_k$ is instantiated using the same textual prediction context as the primary anchor, producing a normalized representation $\boldsymbol{e}_{t_s}$. We then aggregate the similarities across the semantic bank using max pooling Radford et al. (2021); Menon et al. (2020):

$$S(\boldsymbol{x}, c_k) = \max_{s \in \mathcal{S}_k} \operatorname{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{t_s}\right). \tag{4}$$

The resulting score allows each class to be represented by the semantic variant that has the highest affinity with the visual anchor. SER therefore combines taxonomy-aware name expansion with broader text-prototype enrichment, without modifying the visual representation or the evaluated VLM parameters. The semantic bank introduces external text-domain prior knowledge, but its construction uses no target images or image–label pairs.

To limit unnecessary bank-wide matching and reduce over-correction on unambiguous samples during inference, we further introduce an *Adaptive Cascade Inference* strategy. We first compute normalized alignment scores using only the primary labels in $\mathcal{C}$. Let $c_{(1)}$ and $c_{(2)}$ denote the categories with the highest and second-highest primary alignment scores, respectively. The exhaustive semantic rectification is triggered only if the confidence gap between these top candidates falls below a predefined confidence threshold:

$$\begin{aligned} \Delta &= \operatorname{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{c_{(1)}}\right) - \operatorname{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{c_{(2)}}\right), \\ \hat{y} &= \begin{cases} c_{(1)}, & \Delta \geq \tau, \\ \arg\max_{c_k \in \mathcal{C}} S(\boldsymbol{x}, c_k), & \Delta < \tau. \end{cases} \end{aligned} \tag{5}$$

This rule invokes bank-wide matching only when the primary-anchor margin is below τ, reducing unnecessary semantic expansion on higher-margin samples. A low margin is treated as an operational ambiguity signal and is not assumed to identify Taxonomic Bias by itself.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on five standard visual recognition benchmarks: CUB-200-2011 Wah et al. (2011), Oxford Flowers-102 Nilsback & Zisserman (2008), Stanford Cars Krause et al. (2013), Caltech-101 Li et al. (2022a), and ImageNet-1K Russakovsky et al. (2015), covering both fine-grained and general object classification. Additional results on EuroSAT Helber et al. (2017), Food101 Bossard et al. (2014), and DTD Cimpoi et al. (2014) are provided in the Appendix.

Table 1: Zero-shot classification performance across diverse generative VLMs and datasets. Comparison of the reproduced autoregressive baseline Zhang et al. (2024), our Induced Aligned Space method (Ours), and the full Semantic Equivalence Rectification framework (+SER).

Method	Qwen2.5-VL-7B	Qwen3-VL-7B	LLaVA-v1.5-7B	LLaVA-NeXT-7B	Phi-3.5-Vision	DeepSeek-VL2-Tiny	GLM-4.1V-9B	Average
CUB-200								
Base	51.69	46.01	3.92	1.84	7.92	8.65	32.84	21.84
Ours	59.15	46.03	5.84	6.69	11.45	11.77	31.88	24.69 (+2.85)
+SER	60.50	48.31	11.44	11.05	15.79	15.04	39.10	28.75 (+6.91)
Flowers102								
Base	75.04	83.40	9.12	11.01	33.65	38.82	72.80	46.26
Ours	80.47	83.51	22.00	18.93	38.95	39.15	84.62	52.52 (+6.26)
+SER	84.55	84.52	26.06	18.99	42.15	50.01	84.01	55.76 (+9.49)
Cars196								
Base	29.12	44.01	0.05	0.00	3.30	2.47	42.31	17.32
Ours	39.13	44.56	31.94	27.45	22.22	32.92	43.22	34.49 (+17.17)
+SER	45.84	48.32	33.73	25.85	22.31	45.93	51.09	39.01 (+21.69)
Caltech101								
Base	63.63	63.68	46.30	49.41	52.31	49.79	65.76	55.84
Ours	76.15	67.31	74.69	78.01	73.59	77.68	84.66	76.01 (+20.17)
+SER	77.46	79.15	75.99	78.98	75.59	80.73	89.10	79.57 (+23.73)
ImageNet-1K								
Base	41.06	38.83	16.51	17.71	18.45	37.17	38.32	29.72
Ours	57.97	46.56	39.36	38.26	42.30	52.78	53.59	47.26 (+17.54)
+SER	57.58	47.08	39.93	37.60	42.60	53.68	55.03	47.64 (+17.92)

Evaluated Models. We evaluate seven generative VLMs spanning different architectures, model families, and parameter scales: Qwen2.5-VL-7B Bai et al. (2025b), Qwen3-VL-7B Bai et al. (2025a), LLaVA-v1.5-7B Liu et al. (2024a), LLaVA-NeXT-7B Liu et al. (2024b), Phi-3.5-Vision Abdin et al. (2024), DeepSeek-VL2-Tiny Wu et al. (2024), and GLM-4.1V-9B Hong et al. (2025).

Baselines and Proposed Variants. We compare our latent discrimination framework with the closed-world generative classification protocol from prior work Zhang et al. (2024) under a consistent frozen-backbone setting. Specifically, we evaluate the following three settings:

- **Closed-World Generation Baseline Zhang et al. (2024):** We faithfully re-implement the closed-world generative classification protocol from prior work Zhang et al. (2024), where all candidate class names are explicitly enumerated in the prompt and the model is required to generate one class label as the final answer. We use the same closed-world setting because it was shown to substantially outperform open-world generation and therefore constitutes a stronger baseline. To avoid disadvantaging the baseline, we use greedy decoding, deterministic inference, and a label-matching parser that accepts exact matches, case-insensitive matches, and common formatting variants.
- **Induced Aligned Space (Ours):** Our core approach that effectively bypasses explicit textual generation by unifying visual and textual hidden states into a shared comparable metric space, where image features are directly matched against a single primary category anchor.
- **Semantic Equivalence Rectification (SER):** Our full framework that incorporates the augmented semantic bank with max-pooling aggregation and the Adaptive Cascade strategy, thereby explicitly mitigating Taxonomic Bias during generative discrimination.

Evaluation Protocol Scope. All main experiments use frozen generative VLMs without parameter updates or target image–label pairs. The closed-world baseline enumerates the candidate taxonomy in every image prompt, whereas Ours reuses one pre-computed textual anchor per class. SER additionally expands each anchor with externally generated aliases, synonyms, scientific names, and descriptions. Reporting Ours and +SER separately distinguishes the latent-interface gains from those associated with the external text-domain prior.

Table 2: Ablation on the Semantic Bank and Adaptive Cascade mechanism. Accuracy comparison among the baseline Zhang et al. (2024), our latent space method (Ours), indiscriminate semantic bank application (Bank), and the adaptive cascade framework (SER) across seven VLMs on Cars196 and Caltech101.

Model	Cars196				Caltech101			
	Base	Ours	+Bank	+SER	Base	Ours	+Bank	+SER
Qwen2.5-VL-7B	29.12	39.13	38.84	45.84	63.63	76.15	79.55	77.46
Qwen3-VL-7B	44.01	44.56	44.67	48.32	63.68	67.31	66.19	79.15
LLaVA-v1.5-7B	0.05	31.94	29.10	33.73	46.30	74.69	72.00	75.99
LLaVA-NeXT-7B	0.00	27.45	25.28	25.85	49.41	78.01	66.26	78.98
Phi-3.5-Vision	3.30	22.22	18.63	22.31	52.31	73.59	66.15	75.59
DeepSeek-VL2-Tiny	2.47	32.92	28.65	45.93	49.79	77.68	57.83	80.73
GLM-4.1V-9B	42.31	43.22	40.11	51.09	65.76	84.66	85.68	89.10
Average	17.32	34.49	32.18	**39.01**	55.84	76.01	70.52	**79.57**

4.2 Main Results

The main quantitative results on primary benchmarks are summarized in Table 1. At the dataset-average level, our latent methods consistently improve over the reproduced generative baseline across all five evaluated benchmarks.

Performance under Large Candidate Sets. The reproduced closed-world protocol Zhang et al. (2024) is particularly weak for some model–dataset pairs with large candidate sets. For example, on Cars196, LLaVA-v1.5-7B and LLaVA-NeXT-7B obtain 0.05% and 0.00%, while latent matching obtains 31.94% and 27.45%, respectively. On ImageNet-1K, Qwen2.5-VL-7B increases from 41.06% to 57.97%. Because the baseline enumerates all labels in one prompt, these comparisons show the practical advantage of avoiding that specific decoding protocol; they do not by themselves establish that context length is the only cause of the baseline failures.

Taxonomy-Aware Effects of Semantic Rectification. SER further changes the latent decision rule by adding aliases, synonyms, scientific names, and descriptions. It improves several model–dataset pairs, including Qwen2.5-VL-7B on Cars196 from 39.13% to 45.84%, DeepSeek-VL2-Tiny from 32.92% to 45.93%, and Qwen3-VL-7B on Flowers102 from 83.51% to 84.52%.

To examine whether SER benefits specifically from taxonomy–class correspondence, we conduct a controlled sample-level audit with wrong-class mapping controls. The results provide direct evidence that correct taxonomy–class alignment contributes to SER's improvements, while visual descriptions and prototype geometry also play a role. Detailed results are provided in the Appendix.

4.3 Analysis and Discussion

Cross-Model Consistency. Importantly, these classification improvements are not isolated to a specific model family within the evaluated model set. Averaged over seven generative VLMs, our SER framework improves the reproduced baseline from 17.32% to 39.01% on Cars196 and from 55.84% to 79.57% on Caltech101, indicating consistent gains across multiple model architectures.

Q1: Does selective routing improve semantic-bank aggregation?

A1. On average, selective routing performs better than applying the semantic bank to every sample. As shown in Table 2, the naive *Bank* variant degrades the average accuracy compared with our base latent method, dropping from 34.49% to 32.18% on Cars196 and from 76.01% to 70.52% on Caltech101. This indicates that semantic expansion should not be applied indiscriminately.

Table 3 further shows that the effect of τ depends on the dataset and model backbone. On Flowers102, increasing τ improves Qwen2.5-VL from 80.47% to 83.66% and LLaVA-v1.5 from 22.00% to 26.00%. On Caltech101, Qwen2.5-VL peaks at 78.36% with $\tau = 0.010$ and then decreases slightly as more samples are routed to the semantic bank, whereas LLaVA-v1.5 gradually improves

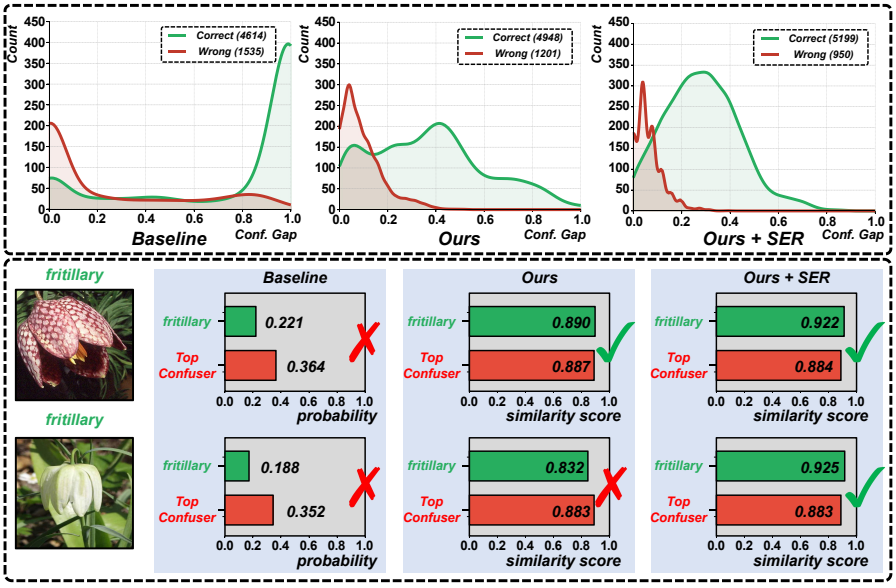

Figure 3: **Relative Confidence Gap analysis and qualitative examples on the Flowers102 dataset. Top:** Normalized Confidence Gap distributions for Baseline, Ours, and Ours+SER; green and red curves denote correct and wrong predictions. **Bottom:** Qualitative case studies where SER enlarges the gap between the ground-truth class and the top confuser, correcting ambiguous predictions.

Table 3: Sensitivity to the adaptive threshold τ on Flowers102 and Caltech101. Accuracy (Acc) and the proportion of samples routed to the semantic bank (Switch) are reported in percentages. At $\tau = 0$, the cascade reduces to primary-anchor latent matching, while larger values trigger semantic rectification more frequently.

Threshold (τ)	Flowers102				Threshold (τ)	Caltech101			
	Qwen2.5-VL		LLaVA-v1.5			Qwen2.5-VL		LLaVA-v1.5	
	Acc	Switch	Acc	Switch		Acc	Switch	Acc	Switch
0.000	80.47	0.00	22.00	0.00	0.000	76.15	0.00	74.69	0.00
0.010	83.29	26.16	23.76	42.22	0.010	78.36	24.75	75.48	18.64
0.020	83.58	42.90	25.07	67.59	0.020	77.68	39.04	75.88	30.06
0.050	83.66	93.74	26.00	95.27	0.050	77.51	78.31	76.38	49.66

from 74.69% to 76.38%. These results show that routing more samples through the semantic bank does not consistently produce larger gains, highlighting the trade-off controlled by τ.

Additional semantic-bank-size ablations are provided in Appendix D.

Q2: Why is selective semantic rectification necessary?

A2. The top panel of Figure 3 shows that *Ours+SER* achieves the clearest separation between correct (green) and wrong (red) predictions. The green curve is sharply peaked at high confidence gaps, indicating that semantic rectification, when applied selectively, effectively amplifies discriminative signals.

More importantly, the bottom case studies in Figure 3 intuitively reveal why blind application harms performance. In Case 1, the *Baseline* misclassifies a *fritillary* as "blackberry lily" (0.3640 vs. 0.2208), but our base *Ours* method already corrects this by boosting the ground-truth score to 0.890. Applying SER here would be redundant and could introduce unnecessary noise. Case 2 tells a different story: while *Ours* significantly improves the ground-truth score from 0.1884 to 0.832, it

still falls short of the confuser "siam tulip" (0.883), resulting in misclassification. This is precisely the type of ambiguous sample that requires semantic rectification. Our full *Ours+SER* successfully resolves this ambiguity by further elevating the ground-truth score to 0.925 (+11.2% over Ours), easily overriding the top confuser.

Although our latent-space formulation provides a contrastive-style interface for generative VLMs, this interface is induced post hoc rather than learned through a native contrastive pre-training objective. Therefore, it is not expected to be as naturally optimized for classification as CLIP. As a complementary capacity check, we train a 16-token soft prompt while keeping all backbone parameters frozen, and apply the same adaptation protocol to CLIP ViT-L/14 as a reference. As shown in Table 4, our latent matching framework substantially improves over the closed-set generative baseline and approaches the CLIP soft-prompt reference on the evaluated benchmarks. This suggests that the proposed interface can effectively expose discriminative information already present in generative VLM representations. Further discussion is provided in the Appendix.

Table 4: Capacity-probing results with soft-prompt adaptation. For generative VLMs, Base denotes the closed-set generative baseline Zhang et al. (2024), and Ours denotes our latent matching framework; for CLIP, Standard denotes the conventional soft-prompt classifier. All backbones are frozen and only the soft prompt is optimized.

Model	Setting	Flowers102	Caltech101	CUB-200
CLIP ViT-L/14	Standard	95.71	96.78	75.27
GLM-4.1V	Base	90.14	89.87	55.50
GLM-4.1V	Ours	98.46	97.57	70.57
Qwen3-VL	Base	91.58	88.74	61.50
Qwen3-VL	Ours	98.05	97.80	69.93

Q3: What are the limitations of this framework in short-context scenarios?

A3. The benefit of our framework can be smaller when advanced VLMs already handle compact candidate sets effectively. As shown in the main results in Table 1, Qwen3-VL-7B improves only marginally from 83.40% to 83.51% with Ours on Flowers102, while +SER reaches 84.52%. This limited improvement suggests that latent matching provides less additional benefit when the backbone already performs strong closed-world discrimination and maps its predictions reliably to the benchmark taxonomy. However, compact candidate sets are not uniformly easy: on Caltech101, +SER improves Qwen3-VL-7B from 63.68% to 79.15%, Qwen2.5-VL-7B from 63.63% to 77.46%, and GLM-4.1V-9B from 65.76% to 89.10%. Although Flowers102 and Caltech101 contain similar numbers of classes, their substantially different gains show that candidate-set size alone is not a reliable indicator of classification difficulty. These results suggest that the effectiveness of our framework also depends on the backbone's closed-world classification ability, the granularity and wording of the candidate taxonomy, and the remaining taxonomy-related label mismatches.

5 Conclusion

In this paper, we studied closed-set image classification with frozen generative VLMs. We proposed a latent matching framework that compares visual and textual hidden states under matched next-token prediction contexts instead of direct label generation. We further introduced Semantic Equivalence Rectification using alternative names, descriptions, and adaptive routing. Experiments across datasets and VLMs show consistent improvements over a reproduced closed-world prompting baseline. A controlled audit shows that taxonomy–class alignment contributes substantially to SER's gains, while semantic and geometric effects also play a role. The gains become smaller when recent VLMs already handle compact candidate sets effectively. Latent representations therefore offer a complementary interface when autoregressive decoding is poorly matched to fixed-label decisions. Future work will investigate target-label-free calibration and representation robustness across prompts, layers, and model families.

AI USE STATEMENT

Generative AI was used in two parts of this work. First, an external large language model was used to generate candidate aliases, scientific names, common synonyms, and visual descriptions for the semantic banks employed by SER. The generated variants were parsed and structured before being used for textual-anchor extraction. Second, a generative-AI assistant was used for format checking and language polishing. The authors reviewed all AI-assisted material.

REPRODUCIBILITY STATEMENT

Section 3 specifies the construction of the visual and textual representations, the similarity-based classifier, Semantic Equivalence Rectification, and the adaptive routing rule. Section 4 describes the evaluated datasets, model backbones, baseline and proposed inference variants, and the main evaluation protocol. The appendix provides additional representation-extraction details, geometric motivation and controlled evidence for hidden-state comparability, auxiliary benchmark results and ablations, the Taxonomic Bias audit, semantic-bank construction details, and clarification of the evaluation scope. These materials are intended to support reproduction and verification of the reported results.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

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.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16388–16398, 2025.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification, 2017.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

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.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, April 2022a.

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, 2022b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024b. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011. URL https://api.semanticscholar.org/CorpusID:16119123.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

APPENDIX

This appendix provides supplementary analyses and implementation details that are not included in the main paper. Specifically, we include: (1) a geometric motivation and controlled empirical evidence for hidden-state comparability under same-token prediction; (2) representation-extraction details; (3) results on three additional visual benchmarks; (4) a Taxonomic Bias audit with controlled semantic-bank interventions and wrong-class mapping permutations; (5) a comparison with a frozen CLIP classifier and a separate supervised soft-prompt capacity probe; (6) semantic-bank construction details and an example; (7) clarification of the evaluation protocol and external text-domain prior; and (8) further discussion of the framework's applicability boundaries.

A GEOMETRIC MOTIVATION AND CONTROLLED EVIDENCE FOR HIDDEN-STATE COMPARABILITY

In this section, we provide a geometric motivation and a controlled empirical probe for the core hypothesis used in our induced aligned space. Contexts that predict the same target token satisfy related logit-margin constraints in the subspace visible to the language modeling head. These shared constraints do not guarantee full-space proximity; we therefore test empirically whether normalized hidden states are sufficiently comparable for similarity-based classification.

Geometric Motivation. Let $h(C) \in \mathbb{R}^d$ denote the last-layer hidden state of a generative language model given an input context C, immediately before generating the next token. The next-token distribution is obtained by projecting this hidden state through the language modeling head:

$$z = Wh(C) + b, \tag{6}$$

where $W \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the output projection matrix, b is the bias term, and $\mathcal{V}$ is the model vocabulary. The probability of generating token y is therefore:

$$p(y \mid C) = \frac{\exp(w_y^\top h(C) + b_y)}{\sum_{j\in\mathcal{V}} \exp(w_j^\top h(C) + b_j)}, \tag{7}$$

where w_y denotes the output-head vector associated with token y.

If a context C predicts token y with high confidence, then the logit of y must be sufficiently larger than the logits of competing tokens. In particular, for any competing token $j \neq y$, a high-confidence prediction of y implies a positive logit margin:

$$(w_y - w_j)^\top h(C) + (b_y - b_j) \gg 0. \tag{8}$$

Equivalently, the hidden state $h(C)$ must lie in a decision region where the projection onto w_y dominates the projections onto other output-head directions. Therefore, if a set of contexts $\{C_i\}_{i=1}^N$ all predict the same token y with high confidence, their hidden states $\{h(C_i)\}_{i=1}^N$ are jointly constrained by similar logit-margin inequalities:

$$(w_y - w_j)^\top h(C_i) + (b_y - b_j) \gg 0, \quad \forall i, \quad \forall j \neq y. \tag{9}$$

These shared constraints place same-token hidden states within a common decision region induced by the language modeling head. However, membership in the same region does not require the hidden states to be close under cosine similarity, since they may retain different context-dependent components.

It is important to note that this does not imply that same-token hidden states must be identical in the full representation space. If there exists a perturbation u such that $Wu = 0$, then:

$$W(h + u) + b = Wh + b. \tag{10}$$

Thus, h and $h + u$ produce exactly the same logits and next-token probabilities, even though they may differ in hidden-state space. More generally, same-token prediction does not mathematically guarantee hidden-state equivalence or cosine proximity. It only imposes related constraints on components visible to the language modeling head. Whether these constraints yield useful similarity structure is an empirical question.

Table 5: Label-level hidden-state similarity matrix. Diagonal entries are within-token similarities, and off-diagonal entries are cross-token similarities.

	airplane	apple	banana	violin
airplane	0.9912	0.7804	0.7451	0.7665
apple	0.7804	0.9790	0.9155	0.7839
banana	0.7451	0.9155	0.9816	0.7832
violin	0.7665	0.7839	0.7832	0.9913

This observation motivates our method rather than proving its validity. We test whether visual and textual contexts associated with the same semantic class produce normalized hidden states that are sufficiently comparable to support metric-based classification without modifying the generative VLM.

Controlled Text-Only Probe. To examine this hypothesis in a controlled setting, we conduct a text-only experiment using `Qwen/Qwen2.5-7B-Instruct`. We use the preset `objects_4way`, where the candidate answer space is restricted to four object tokens:

$$\mathcal{Y} = \{\texttt{banana}, \texttt{airplane}, \texttt{violin}, \texttt{apple}\}. \tag{11}$$

For each target token, we construct eight prompts with different surface contexts but the same intended answer. For each prompt, we extract the last-layer hidden state immediately before the model generates the first answer token. We retain examples whose probability, renormalized over the four candidate tokens, is at least 0.80 and whose top-1 candidate matches the intended target. This candidate-restricted criterion should not be interpreted as confidence over the full vocabulary.

For each retained example, let h_i denote the final-layer hidden state. We first normalize it as:

$$\bar{h}_i = \frac{h_i}{\|h_i\|_2}. \tag{12}$$

We then compute the cosine similarity between every pair of hidden states:

$$\mathrm{sim}(h_i, h_j) = \bar{h}_i^\top \bar{h}_j. \tag{13}$$

Pairs are divided into two groups: same-token pairs, where two hidden states lead to the same predicted token, and cross-token pairs, where they lead to different predicted tokens.

The experiment contains 32 prompts in total, and all 32 examples satisfy the candidate-restricted criterion. Each of the four candidate tokens contributes 8 retained examples. Overall, the mean cosine similarity within the same predicted-token group is 0.9857, whereas the mean cosine similarity across different predicted-token groups is 0.7958, yielding a same-minus-cross similarity gap of 0.1900. The 1-nearest-neighbor label purity in the hidden-state space reaches 0.9688, indicating clear token-conditioned clustering in this controlled setting.

Table 5 shows the label-level hidden-state similarity matrix. The diagonal entries, corresponding to within-token similarities, are consistently higher than the off-diagonal entries, corresponding to cross-token similarities. This provides controlled evidence of token-conditioned clustering before the token is generated.

The matrix exhibits a clear block structure: diagonal values are consistently much higher than most off-diagonal values. The relatively higher similarity between `apple` and `banana` is also reasonable, as both are semantically related fruit categories. Nevertheless, their cross-token similarity remains lower than their respective within-token similarities.

Discussion. These results are consistent with the geometric motivation above. In this controlled text-only setting, hidden states leading to the same next-token prediction are closer on average than hidden states leading to different candidate tokens. The experiment demonstrates an empirical regularity rather than a mathematical consequence of the logit-margin constraints.

This text-only probe does not by itself establish cross-modal hidden-state alignment. Instead, it motivates testing whether the same regularity is useful when visual and textual contexts are compared. The classification experiments in the main paper provide the corresponding task-level evidence for this hypothesis.

B Implementation Details of Representation Extraction

In this section, we clarify how visual and textual representations are extracted from a frozen generative VLM. Our framework does not use generated text as the representation. Instead, for both visual and textual anchors, we extract the hidden state immediately before next-token prediction.

General extraction rule. Given an input context C, we first format it using the model-specific chat template and append the generation prompt. We then perform a single forward pass through the frozen VLM with hidden-state outputs enabled. Let $\boldsymbol{h}_T^{(L)}(C)$ denote the hidden state at the last sequence position T from the final transformer layer L. We define the extracted representation as its ℓ_2-normalized vector:

$$\boldsymbol{e}(C) = \mathrm{Norm}\left(\boldsymbol{h}_T^{(L)}(C)\right) = \frac{\boldsymbol{h}_T^{(L)}(C)}{\left\|\boldsymbol{h}_T^{(L)}(C)\right\|_2}. \tag{14}$$

This position corresponds to the model state immediately before generating the first answer token. Therefore, the extracted representation captures the pre-decoding decision state rather than any generated output.

Visual anchor extraction. For an input image $\boldsymbol{x}$, we construct an image-conditioned prompt $\boldsymbol{t}_x$, such as "Identify the main object in this image." The image and the prompt are jointly fed into the VLM. The visual anchor is extracted as:

$$\boldsymbol{e}_x = \mathrm{Norm}\left(\boldsymbol{h}_T^{(L)}(\boldsymbol{x}, \boldsymbol{t}_x)\right). \tag{15}$$

Thus, $\boldsymbol{e}_x$ is not the output of a standalone visual encoder. It is the final-layer, last-position hidden state of the generative VLM conditioned on both the image and the image query prompt.

Textual anchor extraction. For each category c_k, we construct a textual description prompt $\boldsymbol{t}_{c,k}$, such as "[Description of c_k] is a". This prompt is fed into the same frozen VLM without image input. The corresponding textual anchor is extracted as:

$$\boldsymbol{e}_{t_{c_k}} = \mathrm{Norm}\left(\boldsymbol{h}_T^{(L)}(\emptyset, \boldsymbol{t}_{c,k})\right), \tag{16}$$

where $\emptyset$ denotes the absence of visual input. This representation corresponds to the model state immediately before predicting the class-name token from the textual description.

Similarity computation. After extraction, all visual and textual anchors are compared using cosine similarity:

$$\mathrm{sim}(\boldsymbol{e}_x, \boldsymbol{e}_{t_{c_k}}) = \boldsymbol{e}_x^\top \boldsymbol{e}_{t_{c_k}}, \tag{17}$$

where both vectors are already ℓ_2-normalized. The predicted class is then obtained by nearest-anchor matching:

$$\hat{y} = \arg\max_{c_k \in \mathcal{C}} \mathrm{sim}\left(\boldsymbol{e}_x, \boldsymbol{e}_{t_{c_k}}\right). \tag{18}$$

All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

C Additional Zero-Shot Results on Diverse Benchmarks

We additionally evaluate the framework on Food101 Bossard et al. (2014) (101 food categories), DTD Cimpoi et al. (2014) (47 texture classes), and EuroSAT Helber et al. (2017) (10 land-use/land-cover classes). These benchmarks cover culinary imagery, textural patterns, and satellite remote sensing, extending the evaluation beyond the object-recognition datasets in the main paper.

Results and Analysis. Table 6 reports the results. At the dataset-average level, +SER improves the closed-world baseline from 59.66% to 69.65% on Food101, from 44.21% to 53.67% on DTD, and from 34.66% to 44.11% on EuroSAT. Although the gains vary across backbones, +SER outperforms the baseline on 20 of the 21 evaluated model–dataset pairs; the only exception is Qwen2.5-VL-7B on DTD. This variation is consistent with the boundary discussed in the main paper: performance depends on both the backbone's closed-world classification ability and the structure of the candidate taxonomy.

Table 6: Additional zero-shot closed-set classification accuracy (%) on Food101, DTD, and EuroSAT. Base denotes the reproduced autoregressive baseline Zhang et al. (2024), Ours denotes primary-anchor latent matching, and +SER denotes the full semantic rectification framework with the adaptive cascade.

Model	Food101			DTD			EuroSAT		
	Base	Ours	+SER	Base	Ours	+SER	Base	Ours	+SER
Qwen2.5-VL-7B	84.37	85.56	86.72	61.99	55.85	54.37	29.20	30.77	34.13
LLaVA-v1.5-7B	36.30	53.94	60.03	27.85	41.05	43.07	39.00	58.21	62.42
LLaVA-NeXT-7B	33.10	35.60	44.33	17.30	42.62	44.78	34.57	35.74	37.43
Phi-3.5-Vision	57.61	60.38	66.08	54.32	52.63	55.91	27.07	38.28	38.39
DeepSeek-VL2-Tiny	55.30	61.91	69.30	36.48	53.72	58.61	36.61	47.40	53.89
Qwen3-VL-7B	76.26	77.32	80.26	57.22	60.81	61.55	48.39	44.03	53.19
GLM-4.1V-9B	74.70	74.84	80.82	54.31	57.13	57.39	27.81	29.13	29.33
Average	59.66	64.22	69.65	44.21	51.97	53.67	34.66	40.51	44.11

D Semantic-Bank Size Ablation

Table 7 shows that the reduced-bank settings capture much of the full-bank improvement. In particular, using three semantic variants per class ($M = 3$) approaches the full-bank results across the evaluated settings, while the gains are not uniformly monotonic in M. This suggests that SER does not require exhaustive semantic expansion.

Table 7: Effect of semantic-bank size M under the main evaluation setting. $M = 0$ and $M = \text{All}$ correspond to Ours and +SER, respectively.

M	Flowers102		Cars196	
	Qwen2.5-VL	LLaVA-v1.5	Qwen2.5-VL	LLaVA-v1.5
0	80.47	22.00	39.13	31.94
1	80.56	19.12	44.56	32.98
3	83.64	25.17	45.00	33.47
All	84.55	26.06	45.84	33.73

E Taxonomic Bias Audit

To directly examine whether Semantic Equivalence Rectification (SER) corrects taxonomy-specific label mismatches, we conduct a controlled audit on Flowers102 using a frozen LLaVA-v1.5 model. We use 102 validation images, one per class, to select the SER threshold, and evaluate on 1,020 test images, ten per class. The selected threshold is $\tau = 0.035$. This validation-based choice is specific to the diagnostic audit and is not part of the main zero-shot evaluation protocol. Under this setting, the anchor-only prediction obtains 14.22% accuracy, while full SER improves the accuracy to 20.78%, corresponding to a gain of 6.57 percentage points. At the sample level, SER corrects 113 images that are misclassified by the anchor prediction, while harming 46 originally correct images, leading to a net gain of 67 correct predictions.

We first analyze the 113 SER-corrected samples by testing which subset of the semantic bank can recover each corrected image. The categories are non-exclusive, so one image may be recovered by multiple subsets. As shown in Table 8, 44 out of 113 corrected samples can be recovered by scientific names or aliases, 48 by common synonyms, and 73 by their union. This indicates that taxonomy-related names account for a substantial portion of the corrections. However, visual descriptions also recover 60 samples, suggesting that SER does not operate solely through taxonomic naming. We also observe that unrelated broad expansions recover 59 samples; therefore, this row should not be interpreted as a clean semantic negative control, since adding shared generic candidates can alter the class-wise maximum-similarity scores.

Table 8: Breakdown of SER-corrected samples on Flowers102 with LLaVA-v1.5. The categories are non-exclusive.

Analysis on SER-corrected samples	Ratio
Scientific name / alias	44/113 (38.94%)
Common synonym	48/113 (42.48%)
Visual-description subset	60/113 (53.10%)
Unrelated broad expansion	59/113 (52.21%)
Scientific name / alias ∪ common synonym	73/113 (64.60%)

Table 9: Wrong-class mapping control for taxonomy-related names. Correct taxonomy–class mappings recover substantially more SER-corrected samples than deranged mappings.

Control	Correct mapping	Perm. mean	Perm. max
Scientific name / alias	44	14.02	38
Common synonym	48	5.69	22
Taxonomic union	73	17.83	43

To further test whether these corrections depend on the correct taxonomy–class correspondence, we perform a wrong-class mapping control. Specifically, we keep the same scientific names, aliases, common synonyms, text quantity, and anchor representations, but randomly assign the taxonomy-related names to incorrect classes. We repeat this deranged mapping procedure 200 times. As shown in Table 9, the correctly mapped taxonomy recovers 73 samples, whereas wrong-class mappings recover only 17.83 samples on average. The maximum recovery among all 200 permutations is 43, and no permutation reaches the observed value, yielding a one-sided Monte Carlo p-value of 0.00498.

These results provide direct evidence that Taxonomic Bias is a measurable source of errors: preserving the correct mapping between taxonomy-related names and benchmark classes recovers substantially more samples than assigning the same names to incorrect classes. At the same time, the audit also shows that Taxonomic Bias is not the only factor behind SER's improvement. Visual descriptions and changes in prototype geometry also contribute to the final gain. Therefore, SER should be understood as combining taxonomy-aware rectification with broader semantic prototype enrichment, rather than as a mechanism that exclusively addresses taxonomy-related naming errors.

F Comparison with Frozen CLIP

Table 10 contextualizes the generative-VLM results using a frozen CLIP ViT-L/14 classifier Radford et al. (2021). For each generative backbone, Base is the reproduced closed-world prompting baseline from the main paper, while +SER consistently denotes our latent framework with semantic rectification.

The comparison shows several regimes. On Flowers102, all three +SER rows exceed frozen CLIP, and GLM-4.1V slightly exceeds CLIP on Caltech101. Qwen2.5-VL also comes close to CLIP on CUB-200. These cases suggest that the induced interface can be competitive when the generative backbone contains strong category information and the remaining mismatch can be reduced through latent matching and semantic rectification.

The +SER variants remain below CLIP in most other model–dataset comparisons, with the largest gaps on Stanford Cars and ImageNet-1K. This is expected because CLIP is explicitly trained to calibrate image–text similarity, whereas the proposed space is induced post hoc from hidden states

Table 10: Frozen CLIP and generative-VLM classification accuracy (%). Std. is CLIP's standard frozen zero-shot classifier. Base denotes closed-world generation, while +SER denotes the fixed latent framework with semantic rectification.

Model	Set.	CUB	Flow.	Cars	Cal.	IN-1K
CLIP ViT-L/14	Std.	62.10	74.27	76.20	88.76	74.70
Qwen2.5-VL	Base	51.69	75.04	29.12	63.63	41.06
Qwen2.5-VL	+SER	60.50	84.55	45.84	77.46	57.58
Qwen3-VL	Base	46.01	83.40	44.01	63.68	38.83
Qwen3-VL	+SER	48.31	84.52	48.32	79.15	47.08
GLM-4.1V	Base	32.84	72.80	42.31	65.76	38.32
GLM-4.1V	+SER	39.10	84.01	51.09	89.10	55.03

Table 11: Complete soft-prompt adaptation accuracy (%) on the five main benchmarks. For CLIP, Standard denotes the conventional soft-prompt classifier; for generative VLMs, Base and Ours denote the soft-prompt-adapted closed-world baseline and our latent matching framework, respectively. All backbones are frozen and only the 16-token soft prompt is optimized.

Model	Set.	CUB	Flow.	Cars	Cal.	IN-1K
CLIP ViT-L/14	Standard	75.27	95.71	83.90	96.78	73.82
Qwen2.5-VL	Base	61.27	89.42	65.52	87.45	60.34
Qwen2.5-VL	Ours	68.71	90.08	73.31	95.42	64.31
Qwen3-VL	Base	61.50	91.58	68.74	88.74	61.27
Qwen3-VL	Ours	69.93	98.05	72.87	97.80	66.92
GLM-4.1V	Base	55.50	90.14	70.11	89.87	63.29
GLM-4.1V	Ours	70.57	98.46	75.68	97.57	68.25

optimized for next-token prediction. Fine-grained visual distinctions and large candidate taxonomies can therefore remain difficult even after label-name mismatch is reduced.

Despite this remaining gap, +SER substantially improves the corresponding generative baselines across the reported settings. For example, Qwen2.5-VL improves from 29.12% to 45.84% on Stanford Cars and from 41.06% to 57.58% on ImageNet-1K, while GLM-4.1V improves from 65.76% to 89.10% on Caltech101. These results suggest that internal generative-VLM representations can support structured recognition beyond direct decoding and may motivate related inference interfaces for other fixed-output tasks.

Complete Soft-Prompt Adaptation Results. The main paper reports a compact soft-prompt comparison on three benchmarks. Table 11 provides the complete results on all five main benchmarks using the same Standard, Base, and Ours labels as the main table. This is a supervised capacity probe that uses the labeled training split of each target benchmark and is separate from the main zero-shot evaluation. We optimize a learnable 16-token soft prompt for every setting while keeping all backbone parameters frozen. CLIP uses the conventional soft-prompt classifier, whereas the generative VLMs apply the same lightweight adaptation to the closed-world baseline and our latent matching framework.

Across the three generative backbones, Ours improves over the corresponding Base setting on every reported dataset. Qwen3-VL and GLM-4.1V also exceed the CLIP Standard reference on Flowers102 and Caltech101, while the generative variants remain below CLIP on CUB-200, Stanford Cars, and ImageNet-1K. As a supervised capacity probe, these results indicate that lightweight

prompt adaptation can expose additional discriminative information in the latent interface; they are not included in the zero-shot claims. The remaining gap on fine-grained and large-scale taxonomies also reflects the advantage of CLIP's native contrastive objective.

G Implementation Details of Semantic Bank Construction

To ensure reproducibility and provide transparency regarding our Semantic Equivalence Rectification (SER) framework, we detail the process of constructing the augmented semantic bank.

Generation Process. For a given benchmark, we first extract the complete list of original class names. We then programmatically query Gemini 3 Pro by default for each class. The LLM is instructed to act as an expert taxonomist and provide a list of top M semantic equivalents specific to the dataset's visual domain. We strictly constrain the LLM to generate highly accurate descriptive synonyms, formal scientific nomenclature, and common aliases that represent the exact same category, explicitly forbidding the generation of broader parent categories (hypernyms). Finally, we apply a lightweight post-processing script to parse the LLM's response, strip any conversational fillers, and structure the variants into a clean mapping dictionary for subsequent textual hidden state extraction.

Concrete Example. To illustrate the quality and necessity of these generated semantic variants, we provide a concrete example from the Oxford Flowers-102 dataset:

- **Original Label:** "fritillary"
- **Generated Semantic Bank ($M = 3$):** ["Fritillaria", "Snake's head fritillary", "Fritillaria meleagris"]
- **Analysis:** The variants cover a Latin genus, a common name, and a species name, illustrating the types of alternative expressions included in the semantic bank.

H Evaluation Protocol Scope

In the main zero-shot experiments, excluding the supervised soft-prompt capacity probe above, all evaluated generative VLMs remain frozen, and neither Ours nor +SER uses parameter updates or target image–label pairs. The primary latent-matching method constructs one textual anchor from each benchmark label. SER additionally uses externally generated aliases, synonyms, scientific names, and descriptions, introducing an external text-domain prior without modifying the evaluated VLM.

Accordingly, the main paper reports Ours and +SER separately. The comparison with the closed-world baseline should be interpreted as frozen-backbone inference under different textual interfaces: the baseline enumerates the candidate labels in each image prompt, Ours pre-computes one textual anchor per class, and +SER expands those anchors with external textual variants. This protocol isolates the effect of the inference interface while making the additional information used by SER explicit.

The soft-prompt capacity probe follows a different protocol: it uses labeled training images from each target benchmark to optimize only the 16 prompt vectors. It is reported solely to examine accessible classification capacity and is not part of the zero-shot evaluation.

I Further Discussion on Applicability Boundaries

The main paper shows that the benefit of latent matching can shrink when recent VLMs already handle compact candidate lists effectively. This boundary should not be attributed to class count alone. It also depends on the backbone's instruction-following ability, the granularity and wording of the candidate labels, and how well the visual concepts are represented by the benchmark taxonomy.

For short-context settings with compact taxonomies, closed-world prompting can condition each prediction on the complete label list and exploit candidate-specific interactions that are absent from

a generic visual anchor. As generative VLMs improve, this can narrow the advantage of latent matching or make the Base setting stronger for some model–dataset pairs. In contrast, long or fine-grained candidate lists increase repeated prompt processing and can expose label-name mismatches, making cached textual anchors and semantic variants more useful. Our results do not isolate context length as the sole cause; instead, they indicate an interaction among backbone capability, taxonomy structure, and inference interface.

The frozen CLIP comparison provides a complementary boundary. Contrastive pre-training remains advantageous when globally calibrated image–text distances and fine-grained visual separation are central, particularly on Stanford Cars and ImageNet-1K. Latent matching is therefore best viewed as an adaptation for reusing generative backbones rather than as a replacement for natively contrastive classifiers. Its practical value is strongest when a deployed generative VLM must also support fixed-output recognition without an additional visual model, or when textual anchors can be pre-computed and reused across images.

Future work should examine target-label-free routing, robustness across prompts and layers, and hybrid interfaces that select among autoregressive generation, latent matching, and contrastive retrieval according to candidate-set structure and model confidence.