RETHINKING LONG-TAIL LEARNING IN VISION-LANGUAGE MODELS: FROM SINGLE DISTRIBUTION TO MODALITY-COUPLED DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have achieved remarkable progress in multimodal processing and generation, yet still exhibit systematic failures in object hallucination and rare-concept understanding. Prior studies have shown that the concept distribution in multimodal data is markedly long-tailed, resulting in imbalanced downstream performance. However, existing long-tail methods mainly employ a single frequency statistic to define long-tail distribution, which cannot accurately reflect the effective long-tail distribution in multimodal scenarios. For example, the same concept may exhibit markedly different scarcity across modalities, leaving some concepts are frequent in text but weakly grounded visually. Moreover, shared multimodal representation spaces entangle long-tail biases across modalities, making it difficult to decouple modality-specific effects, causing incomplete rare-concept expression. We refer to this problem as modality-inconsistent long-tail. To address this, we propose a pre-decoupled multimodal-aware long-tail learning framework (PMLL), which introduces modality-decoupled reweighting to model multimodal long-tail distributions, rare-word reinforcement to improve rare token expression, and adaptive task-aware optimization to balance multi-task contribution. Furthermore, a keyword-decoupled long-tail evaluation metric is introduced to discover the detailed learning performance of informative tokens, while a metric-general co-occurrence-aware long-tail mining protocol is developed to reduce co-occurrence-induced bias in estimating long-tail distributions. Overall, we provide a unified framework for understanding, mitigating, and evaluating modality-inconsistent long-tail effects in VLM learning. Extensive experiments demonstrate the effectiveness of the PMLL under modality-inconsistent long-tail.

1 INTRODUCTION

Recent advances in artificial intelligence have driven rapid progress in vision-language models (VLMs) (Li et al., 2025), enabling increasingly strong performance across a wide range of multimodal tasks, including image captioning (Xing et al., 2026), visual question answering (An et al., 2026), and visual recognition (Ortego et al., 2026). However, real-world multimodal data are inherently long-tailed (Xu et al., 2026), where a large number of rare concepts receive limited supervision, posing a fundamental challenge to VLM learning. Moreover, tail concepts are not only underrepresented in the training data, but are also frequently correlated with or co-occur with dominant head concepts (Zhong et al., 2026), causing their already limited supervision to be further overwhelmed by head concepts knowledge. As a result, VLMs tend to exhibit degraded performance, leading to problems such as incorrectly generated tail concepts, object hallucination (Huang et al., 2025), and systematic bias toward frequent head concepts, which can be particularly severe in safety-critical applications. Recent studies have begun to address this issue by estimating concept-frequency distributions in VLM training data and using such statistics to identify and mitigate underrepresented concepts. However, most existing approaches characterize the long-tail structure through a single aggregated distribution (Chen et al., 2025), implicitly assuming that long-tail are consistent across modalities.

This is insufficient for realistic multimodal data, where the amount and quality of learnable evidence associated with a concept can differ substantially between the visual and textual modalities. We refer to this more complex long-tail phenomenon, jointly determined by concept imbalance and

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.

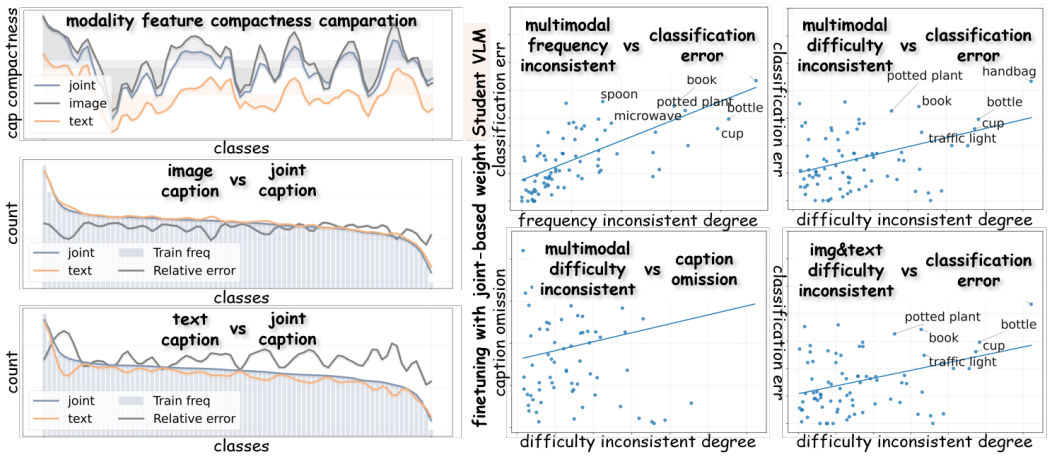


Figure 1: Analysis of long-tail issues in multimodal VLM fine-tuning. "Joint" denotes simultaneous image-text input, with difficulty measured by frequency and compactness. We compare modality-specific long-tail distributions and performance variations with modality-inconsistent long-tail degree.

inconsistency across modalities, as the **modality-inconsistent long-tail**. Since VLMs are commonly applied to multiple downstream tasks, we examine modality-inconsistent long-tail effects across both captioning and classification tasks through reweighted fine-tuning experiments on COCO (Chen et al., 2015) with QwenVL (Wang et al., 2024a), as illustrated in Fig. 1. Specifically, reweighting follows the aggregated image-text distribution assumption adopted by previous works (Wang et al., 2024b), using joint image-text inputs to derive long-tail weights, while the modality-inconsistent long-tail degree is characterized by the discrepancy among image-only, text-only, and joint distributions in class frequency and feature compactness. The results show that **more severe performance degradation emerges along with increasing modality-inconsistent long-tail degree**. Moreover, the unified representation space in multimodal models further entangles the modality-inconsistent long-tail effects, resulting in more complex long-tail patterns that can lead to severe consequences. For example, Jain et al. (2021) show that abnormalities such as pulmonary nodules may be visually observable in medical images but absent from corresponding reports, resulting in insufficient textual supervision for rare concepts and biased model learning, leading to missed diagnoses.

Overall, modality-inconsistent long-tail distributions hinder VLM learning and causing inaccurate predictions, hallucinations, and tail concept underrepresentation. To address this, we develop a **pre-decoupled multimodal-aware long-tail learning framework (PMLL)** by jointly modeling long-tail distributions from global, visual, and textual perspectives. Specifically, PMLL introduces a modality-decoupled long-tail re-weighting module for modeling and decoupling cross-modal long-tail discrepancies, a rare word reinforcement mechanism for improving the diversity and consistency of model generation, and an adaptive task-aware optimization modulation strategy for mitigating multi-task optimization conflicts. Beyond training, tail concepts are often coupled with dominant head concepts through class co-occurrence, making their actual long-tail characteristics difficult to identify. Moreover, existing metrics are not specifically designed to evaluate long-tail learning of informative tokens, leading to insufficient emphasis on informative concepts and obscured long-tail effects in evaluation results. To address this, a co-occurrence-aware long-tail mining protocol and a keyword-decoupled long-tail evaluation metric are introduced to more faithfully reveal the model's capability in learning and expressing rare concepts. Our main contributions are:

- We reveal that existing multimodal long-tail methods overlook modality-specific long-tail distributions by assuming a single aggregated one, which we define as modality-inconsistent long-tail, and propose PMLL to model coupled modality-specific distributions.
- To alleviate co-occurrence bias and informative token underemphasis, we introduce a co-occurrence-aware protocol and keyword-decoupled metric for reliable long-tail evaluation.
- Extensive experiments across captioning, classification generation, and VQA tasks validate the effectiveness of PMLL in improving multimodal long-tail understanding.

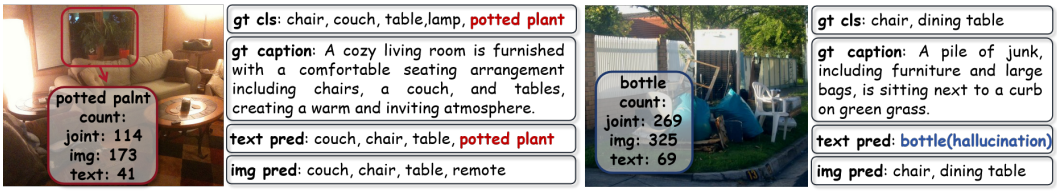


Figure 2: Examples of modality-inconsistent long-tail effects under fine-tuning with long-tail weights derived from image or text-only inputs. "Potted plant" is under-reweighted and prone to omission, while "bottle" is over-reweighted and more likely to hallucinate due to inaccurate long-tail weighting.

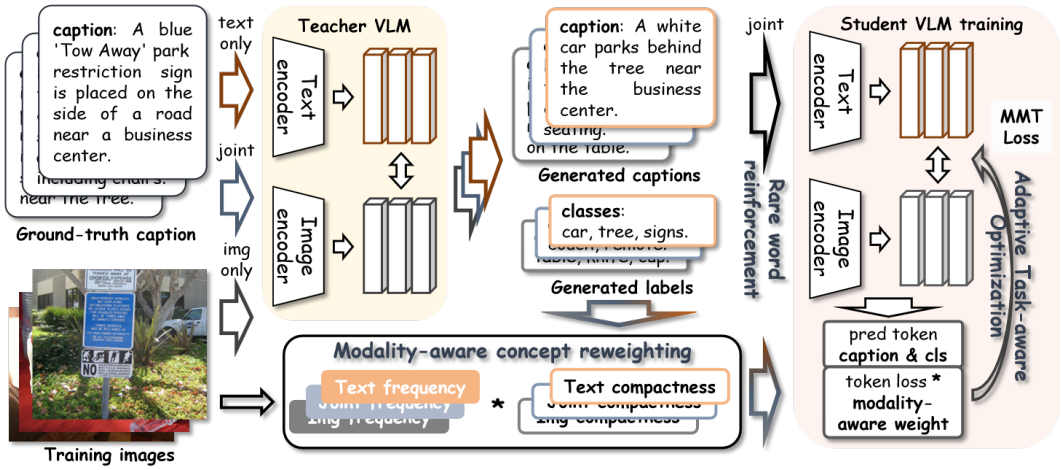


Figure 3: Overview of PMLL with three key components, including modality-aware concept reweighting, rare word reinforcement, and adaptive task-aware optimization.

2 METHOD

2.1 PROBLEM DEFINITION

Let the training set be $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i, \mathbf{y}_i)\}_{i=1}^{N}$,where $\mathbf{x}_i$ denotes an image, $\mathbf{t}_i$ denotes its paired text or instruction, and $\mathbf{y}_i$ denotes the target response sequence. In the unified generative formulation, the response can contain both caption and classification supervision. Conventional long-tailed learning method characterizes the rarity of a concept c using its multimodal-aggregated occurrence frequency n_c, and derives a reweighting coefficient as $w_c = f_r(n_c)$. However, such an aggregated statistic can be insufficient for multimodal learning because it implicitly assumes consistent modality long-tailed distributions. Examples with large cross-modal long-tail gaps are shown in Fig. 2, demonstrating the failure of conventional statistics in capturing modality-specific evidence. To formalize this observation, we consider three observation views, $m \in \{V, T, J\}$ for a concept c, corresponding to the visual-only, text-only, and joint image-text settings with observed frequencies and compactness as n_c^m, q_c^m to characterize the amount and quality of the learned concept representations, respectively. A larger q_c^m suggests a more stable and discriminative representation.

2.2 PRE-DECOUPLED MULTIMODAL-AWARE LONG-TAIL LEARNING

With multiple observation views, we propose PMLL (**P**re-decoupled **M**ultimodal-aware **L**ong-tail **L**earning) framework to address the modality-inconsistent long-tail, which consist of three key components as shown in Figure 3. First, **Modality-decoupled Long-tail Re-weighting** is constructed by jointly considering the concept frequency and representation compactness of different viewpoints. Second, since caption words beyond category concepts also exhibit long-tailed distributions, **Rare Word Reinforcement** is introduced to enrich fine-grained descriptions and improve rare concept expression in captions. Moreover, due to the varying difficulties and objectives across tasks, **Adaptive Task-aware Optimization** is introduced to dynamically balance their contributions during training.

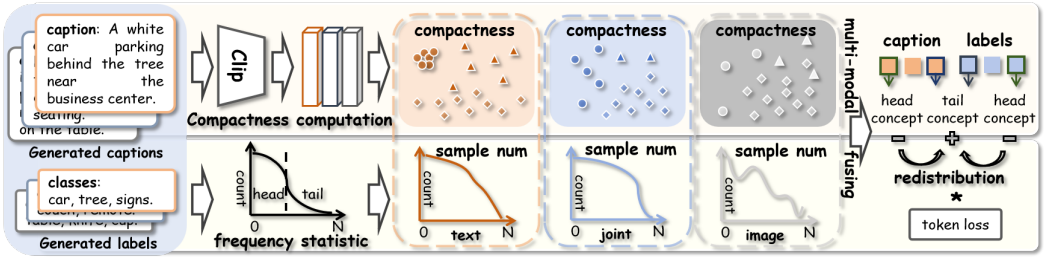


Figure 4: Overview of modality-inconsistent concept mining and re-weighting.

2.2.1 Modality-decoupled Long-tail Reweighting

Modality-decoupled long-tail mining. To mine the effective long-tail patterns of different modalities, a stronger teacher VLM is first employed as a modality-aware statistical estimator to construct concept-level statistics from three complementary observation views for training set D_t as shown in Fig. 4. For a training sample i, the teacher performs visual-only, text-only, and joint image-text concept extraction, producing $\mathbf{y}_i^V, \mathbf{y}_i^T, \mathbf{y}_i^J$. After that, to reveal the complex coupled long-tail distribution for each modality, the occurrence frequency n_c^m and compactness q_c^m of each concept $c \in C$ is extracted from $\mathbf{y}^m$. Specifically, q_c^m is computed using representation $z_{i,c}^m$ extracted from CLIP by

$$q_c^m = \frac{1}{\epsilon + \frac{1}{|\mathcal{I}_c^m|}\sum_{i\in\mathcal{I}_c^m}\left\|z_{i,c}^m - \mu_c^m\right\|_2^2}, \tag{1}$$

where $\mu_c^m = \frac{1}{|\mathcal{I}_c^m|}\sum_{i\in\mathcal{I}_c^m} z_{i,c}^m$ is the feature centroid of c. Together, n_c^m and q_c^m approximate the modality long-tail characteristics by jointly considering the quantity and quality of concept feature.

Modality-decoupled long-tail weighting. Specifically, the modality-specific long-tail is quantified from both quantity and representation quality for reweighted long-tail learning by

$$\tilde{w}_{c,m}^n = \sqrt{\frac{N^m}{n_c^m + \epsilon}}, \qquad \tilde{w}_{c,m}^q = \sqrt{\frac{\bar{q}^m}{q_c^m + \epsilon}}, \qquad w_{c,m} = \frac{\tilde{w}_{c,m}}{\frac{1}{|\mathcal{C}_m|}\sum_{k\in\mathcal{C}_m}\tilde{w}_{k,m}}, \tag{2}$$

where $N^m = \sum_{c\in C} n_c^m$ denotes the total concept occurrences in modality m, $\bar{q}^m = \frac{1}{|\mathcal{C}_m|}\sum_{c\in C} q_c^m$ denotes the mean concept compactness. $\tilde{w}_{c,m}^n$ and $\tilde{w}_{c,m}^q$ are normalized for scale stability. The final modality-decoupled weight for concept c is obtained by integrating the factors across modalities as

$$w_c = (\lambda_V w_{c,V}^n + \lambda_T w_{c,T}^n + \lambda_J w_{c,J}^n)(\lambda_V w_{c,V}^q + \lambda_T w_{c,T}^q + \lambda_J w_{c,J}^q), \tag{3}$$

where λ_V, λ_T, and λ_J control the contributions of the visual, textual, and joint views, respectively. By doing this, the resulting weight better reflects the effective long-tail distribution under modality inconsistency, thereby alleviating modality-inconsistent long-tail effects.

Moreover, to further alleviate the excessive influence of dominant head concepts on tail concepts within the same sample, a tail compensation term is introduced for tail-related class tokens. Specifically, the final token weight for concept c is formulated as

$$\omega_{i,k}^{final} = 1 + \eta_{r(i,k)}\left(\omega_{i,k} + d_i \mathbb{I}[c(i,k) \in Tail] - 1\right), \quad d_i = \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\max(1 - w_c, 0), \tag{4}$$

where $\mathcal{H}_i = \{c \in \mathcal{H} | c \in C_i\}$ denotes the set of head concepts appearing in sample i. $\eta_{r(i,k)}$ is a residual scaling coefficient for different response regions. Specifically, since classification supervision directly reflects concept learning with shorter sequences, a larger coefficient is assigned to classification tokens ($\eta_{\text{cls}} = 0.5$) than caption tokens ($\eta_{\text{cap}} = 0.1$) to balance their contributions. By doing this, the VLM can better capture tail knowledge in head dominated sample.

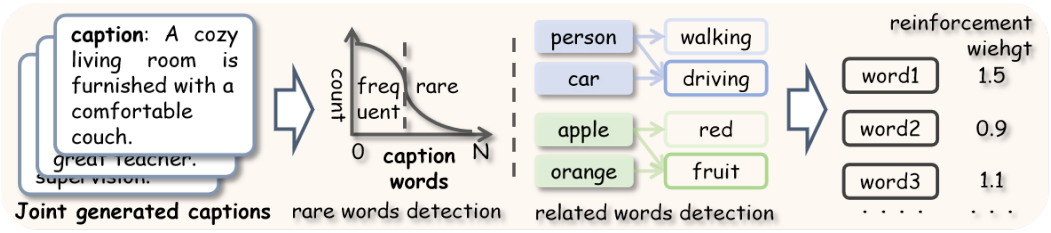


Figure 5: Overview of rare word reinforcement mechanism.

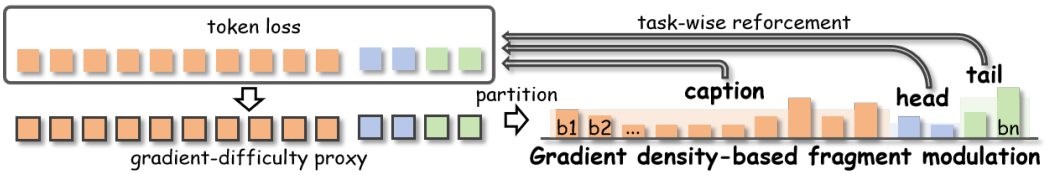


Figure 6: Overview of adaptive task-aware optimization.

2.2.2 Rare Word Reinforcement

Modality-decoupled long-tail reweighting mainly strengthens explicit concept tokens, while leaving rare words, including tail-related attributes and low-frequency words, insufficiently reinforced. For this reason, a rare-word reinforcement mechanism is introduced to strengthen the supervision of rare words, as shown in Fig. 5. Specifically, rare words mined from the training caption set are enforced according to their related concepts c and frequency n by

$$\omega_{i,k}^{rel} = 1 + \rho(w_c - 1), \qquad \omega_{i,k}^{low} = 1 + \rho\left(N^J / n_{rare}^J\right), \tag{5}$$

where ρ is set to 0.5 empirically. Specially, if the same attribute is related to multiple concepts, the maximum w_c will be chosen. By doing this, the tail words beyond explicit concepts can be reinforced, leading to more fine-grained and diverse caption generation.

2.2.3 Adaptive Task-aware Optimization

Concept long-tail and cross-modal inconsistency determine what information should receive additional supervision. However, different task-related textual segments can still exhibit substantially different and difficult-to-quantify learning difficulties during multi-task generative training. To address this, inspired by GHM (Li et al., 2019), adaptive task-aware optimization is designed to model task-specific learning status for the overall objective as shown in Fig. 6. First, to estimate the current token difficulty states, the gradient-difficulty proxy $g_{i,k} = 1 - p_{i,k}$ is computed based on the the standard token-level autoregressive cross-entropy loss $\ell_{i,k}$ and the target probability $p_{i,k} = \exp(-\ell_{i,k})$, where a larger $g_{i,k}$ indicates a lower degree of learning. Then, to locate the rare token range, tokens are partitioned into $b \in \{1, \ldots, B\}$ intervals according to $g_{i,k}$ and the group scarcity is defined as

$$g_{i,k}^{gro} = \frac{\bar{D}}{D_{b(i,k)} + \epsilon}, \quad D_b = \mu D_b^{(t-1)} + (1-\mu) D_b^{(t)}, \tag{6}$$

where D_b denotes the estimated density of group b, which is updated using historical density states to obtain a stable estimate during training. $\bar{D}$ denotes the average density over valid intervals, and μ is the momentum coefficient. A larger $g_{i,k}^{gro}$ indicates a relatively sparse region.

After obtaining dynamically estimated token-level group scarcity, it is aggregated over task segments $S = \text{cap}, \text{head}, \text{tail}$ by $\frac{1}{|\Omega_s|} \sum_{(i,k)\in\Omega_s} g_{i,k}^{gro}$ to estimate segment-level optimization difficulty, where the three segments correspond to caption generation, head-concept supervision, and tail-concept supervision, respectively. Ω_s denotes valid tokens in segment s. The coefficient for segment s is

then defined as $\lambda_s = \frac{\bar{g}_s^{gro}}{\bar{g}^{gro}+\epsilon}$, where $\bar{g}^{gro}$ is the global mean group scarcity. Finally, the weighting objective can be defined as

$$\mathcal{L}_{PMLL} = \frac{1}{Z}\sum_{s\in S}\lambda_s\mathcal{L}_s, \qquad \mathcal{L}_s = \sum_{(i,k)\in\Omega_s}\omega_{i,k}^{rel}\omega_{i,k}^{low}\omega_{i,k}^{final}\ell_{i,k}, \tag{7}$$

where $Z = \sum_{s\in S}\lambda_s|\Omega_s|$ is used for normalization. In this way, the relative contributions of task-related supervisions can be adaptively balanced according to their current optimization states. Overall, PMLL jointly optimizes these task-related regions with adaptive strengths guided by word rarity, cross-modal long-tail inconsistency, and segment-level difficulty.

3 Related Work: Long-tail VLMs Learning

Real-world multimodal datasets naturally exhibit long-tailed distributions, which further aggravate tail-class underrepresentation and hallucination in VLMs (Parashar et al., 2024). Considerable efforts have been devoted to addressing this issue. For example, Ma et al. (2022) leverage pretrained vision-language representations and balanced adaptation to improve long-tailed visual recognition. Wang et al. (2024b) further investigate how conventional imbalanced-learning strategies, including Focal Loss and Balanced Softmax, can be integrated with pretrained VLMs. Ye et al. (2025) exploits a large VLM to extract semantic keywords and identify rare samples according to keyword frequency. However, they implicitly assume a single long-tail distribution for multimodal data, which overlooks the inherent inconsistencies across modalities as well as the missing cross-modal evidence. Song et al. (2025) generate additional image-text data to more precisely supplement globally scarce tail concepts, yet the multimodal long-tail distribution is still essentially treated as a unified global distribution. In contrast, our method explicitly decouples modality-inconsistent long-tail distributions, thereby approximating a more realistic and informative characterization of long-tail patterns.

4 Experiments

4.1 Experimental Setup

Datasets. The experiments are conducted on three widely used multimodal benchmarks, including MS COCO 2017 (Chen et al., 2015), Conceptual Captions 3M (CC3M) (Sharma et al., 2018), and Visual Genome (VG) (Krishna et al., 2017). MS COCO 2017 serves as the primary benchmark for image captioning, providing well-established image-text pairs for analyzing category-level long-tail distributions. CC3M contains approximately 3.3M image-text pairs with more diverse and realistic descriptions, enabling evaluation on large-scale real-world multimodal distributions with naturally occurring long-tail concepts. Visual Genome further extends the evaluation with rich region-level annotations, and visual question answering annotations. Specially, for consistent cross-dataset analysis, 80 COCO categories serve as a shared concept space for all experiments.

Evaluation Metrics. For classification, multi-label accuracy is reported to assess concept recognition in images. For evaluating caption generation, standard captioning metrics, including BLEU, CIDEr, ROUGE-L (R-L), and SPICE, are adopted. Besides text-level similarity, several model-based cross-modal evaluation metrics are further employed. Specifically, CLIP-S (C-S) and REF-CLIP-S (R-C-S) (Hessel et al., 2021) evaluate the semantic alignment between captions and visual contents, while PAC-S++ (P-S++) and REF-PAC-S (R-P-S++)++ (Sarto et al., 2025) measure image-text compatibility from a learned cross-modal representation space. Polos (Wada et al., 2024) is additionally adopted to provide a complementary evaluation of generation quality.

Furthermore, conventional captioning metrics mainly measure overall caption-reference similarity, overlooking informative concepts and associated words, which limits their effectiveness in evaluating long-tail learning. For this reason, we introduce a general principle that is not tied to any specific text- or feature-level metric, named **Keyword-decoupled Long-tail Evaluation (KLE)**, which provides a more informative approximation of effective long-tail imbalance by emphasizing key words and semantic consistency. Specifically, each token t is assigned a semantic weight $\omega_t \in \{w_{cls}, w_{rel}, w_{con}, w_{stop}\}$ according to its semantic role derived from the training corpus, where

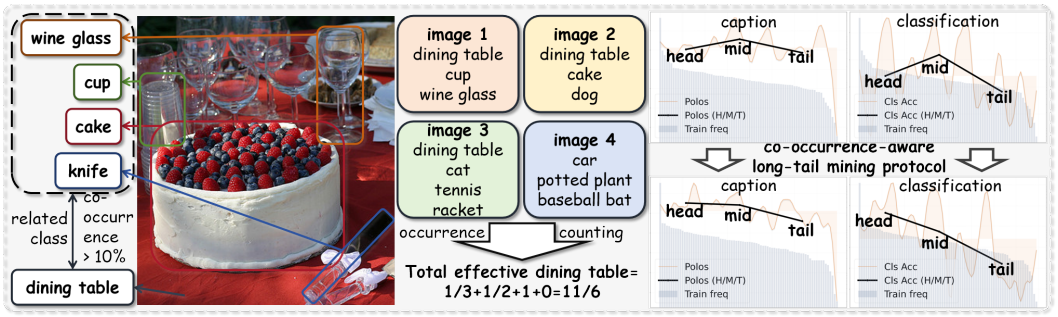


Figure 7: Illustration of co-occurrence-aware tail mining. The proposed protocol better reveals the underlying long-tail effects after correcting co-occurrence-induced frequency bias.

category concepts, concept-related words, normal content words, and stop words are assigned different importance levels. For concept-related words, w_{rel} is further adjusted according to their semantic relevance and distance-based association with their corresponding concept. In our implementation, these semantic weights are incorporated into CIDEr to measure long-tail-aware textual similarity. Furthermore, feature-level consistency among the image, generated caption, and ground-truth caption is additionally quantified using weighted CLIP-S and Ref-CLIP-S. By jointly emphasizing key words and multimodal semantic consistency, KLE provides a more faithful assessment of long-tail learning.

Moreover, for the above caption and classification metrics, frequent co-occurrence between tail and head concepts can inflate the nominal frequency of tail concepts despite limited independent supervision, making count-based tail identification unreliable. To address this, a **metric-agnostic co-occurrence-aware long-tail mining protocol** is introduced to identify a more realistic long-tail distribution, as shown in Fig. 7. Specifically, let $\mathcal{C}_i$ denote the set of concepts appearing in sample i, and $\mathcal{N}(c)$ denote the set of concepts frequently co-occurring with concept c. A concept c' is included in $\mathcal{N}(c)$ when its co-occurrence frequency with c exceeds 10% of n_c. The co-occurrence-adjusted rarity of concept c in sample i is defined as $r_{i,c} = \frac{1}{1+|\mathcal{C}_i \cap \mathcal{N}(c)|}$, where $|\cdot|$ denotes the set cardinality. A larger $r_{i,c}$ indicates that concept c receives more independent supervision with fewer co-occurring concepts. Importantly, this protocol provides only a coarse prior of long-tail characteristics rather than a direct estimation of effective learning evidence, and is therefore used solely for evaluation.

Baselines. PMLL is compared with several representative long-tail learning baselines, including VLM-adapted Focal Loss and Balanced Softmax (BS) (Wang et al., 2024b), Compensating Logit Adjustment (CLA) (Shi et al., 2024), LinUpper (Sow et al., 2025), TIMER-UL (TUL) (Wu et al., 2023), GALA (Zhang et al., 2024), Token Reweighting (TR) (Weers et al., 2026), and Bounded Adaptive Loss (BAL) (Miao et al., 2026). Focal Loss emphasizes difficult tokens, while BS and CLA correct frequency-induced bias. LinUpper performs online difficulty-based reweighting, TUL suppresses frequent-token generation, and GALA adaptively calibrates predictions using online token statistics. BAL adaptively controls high- and low-confidence token contributions to alleviate head-class overfitting. Specially, TR is adapted to the long-tailed setting by replacing its uniform token weight with class-specific imbalance-aware weights.

4.2 Main Results

To evaluate PMLL under modality-inconsistent long-tail distributions, as shown in Tab. 1, we conduct VLM fine-tuning on COCO, CC3M, and VG using Qwen2.5-32B and Qwen2.5-3B (Bai et al., 2025) as teacher and student models, respectively. Only 10% of each dataset is randomly sampled for fine-tuning. Moreover, cross-architecture robustness evaluation are conducted using Qwen3.5-4B (Qwen Team, 2026) and InternVL3.5-4B (Wang et al., 2025) as students as shown in Tab. 2. It shows that PMLL achieves the best overall performance under joint optimization of captioning and classification and remains stable on the more challenging CC3M and VG datasets, where several baselines degrade severely. Although some methods such as BAL perform better on individual tasks, their performance drops sharply on others, confirming the stronger multi-task robustness of PMLL. Moreover, KLE enables robust cross-dataset discrimination and word-type-specific caption evaluation, with additional results using Gemma 4-26B (Gemma Team, 2026) as the teacher provided in the Appendix A.2.

Table 1: VLM fine-tuning results for joint captioning and classification. Best results are in bold.

Task	Captioning										Classification				Dataset
Metric /Method	BLEU	CIDEr	R-L	SPICE	C-S	R-C -S	P-S++	R-P -S++	Polos	KLE	Head	Mid	Tail	Avg	
Softmax	0.476	3.368	0.621	0.566	0.792	0.845	**0.580**	**0.708**	0.782	4.350	0.861	0.797	0.753	0.799	**COCO (Overall avg relative increase for PMLL 0.1)**
Focal	0.441	3.066	0.594	0.538	0.791	0.843	0.579	0.706	0.775	4.053	0.851	0.775	0.711	0.772	
CLA	0.452	3.122	0.604	0.546	0.792	0.843	0.579	0.705	0.775	4.082	0.862	0.791	0.748	0.795	
BS	0.434	2.934	0.591	0.532	0.791	0.841	0.578	0.704	0.768	3.931	0.845	0.790	0.756	0.793	
LinUpper	0.469	3.306	0.615	0.561	0.791	0.844	0.578	0.706	0.779	4.267	0.854	0.798	0.751	0.796	
TUL	0.477	3.386	0.622	0.567	0.792	**0.846**	**0.580**	**0.708**	**0.783**	4.364	0.860	0.797	0.752	0.798	
Gala	0.476	3.372	0.621	0.566	0.792	0.845	**0.580**	**0.708**	0.782	4.353	0.861	0.797	0.752	0.798	
TR	0.468	3.294	0.615	0.559	0.792	0.845	0.578	0.707	0.780	4.164	0.846	0.793	0.753	0.793	
Bal	0.478	3.401	**0.623**	**0.568**	0.792	**0.846**	**0.580**	**0.708**	**0.783**	4.369	**0.862**	**0.799**	0.756	0.801	
PMLL	**0.479**	**3.406**	**0.623**	**0.568**	**0.793**	**0.846**	**0.580**	**0.708**	**0.783**	**4.375**	0.859	0.796	**0.765**	**0.802**	
Softmax	0.457	3.335	0.595	**0.547**	0.782	0.843	0.595	0.722	0.778	3.771	0.766	0.703	0.586	0.675	**CC3M (Overall avg relative increase for PMLL 0.4)**
Focal	0.417	2.955	0.565	0.508	0.785	0.842	**0.596**	0.721	0.770	3.512	0.729	0.670	0.522	0.628	
CLA	0.437	3.104	0.578	0.514	0.784	0.841	**0.596**	0.721	0.770	3.425	0.756	0.630	0.490	0.612	
BS	0.385	2.606	0.536	0.461	0.781	0.833	0.593	0.714	0.746	2.871	0.724	0.708	0.675	0.700	
LinUpper	0.451	3.297	0.591	0.538	0.783	0.843	0.595	0.722	0.776	3.740	0.763	0.707	0.597	0.680	
TUL	0.456	3.322	0.594	0.547	0.783	0.843	0.595	0.722	0.778	3.777	0.766	0.710	0.594	0.680	
Gala	0.360	2.514	0.517	0.468	0.781	0.831	0.590	0.711	0.738	3.225	0.665	0.603	0.472	0.569	
TR	0.455	3.319	0.595	0.544	**0.784**	**0.844**	**0.596**	**0.723**	0.778	3.756	0.761	0.733	**0.664**	0.714	
Bal	**0.459**	3.357	**0.597**	**0.548**	0.783	0.843	0.595	0.722	0.778	**3.797**	**0.768**	0.707	0.594	0.680	
PMLL	0.458	**3.368**	0.596	0.547	0.782	**0.844**	**0.596**	**0.723**	**0.779**	3.775	0.765	**0.735**	0.661	**0.715**	
Softmax	0.381	1.715	0.487	0.452	**0.793**	**0.838**	**0.577**	**0.699**	0.796	3.358	**0.730**	0.635	0.465	0.595	**VG (Overall avg relative increase for PMLL 0.2)**
Focal	0.355	1.541	0.464	0.435	0.789	0.834	0.576	0.697	0.789	3.171	0.674	0.568	0.405	0.534	
CLA	0.356	1.499	0.465	0.430	0.790	0.833	**0.577**	0.697	0.788	3.083	**0.730**	0.626	0.446	0.585	
BS	0.001	0.007	0.002	0.063	0.643	0.614	0.483	0.507	0.220	0.777	0.684	0.590	0.430	0.554	
LinUpper	0.363	1.572	0.470	0.436	0.792	0.835	0.576	0.696	0.790	3.186	**0.730**	0.628	0.456	0.590	
TUL	0.382	1.720	0.487	0.452	0.792	**0.838**	**0.577**	**0.699**	0.796	3.356	0.728	0.632	0.460	0.592	
Gala	0.369	1.615	0.476	0.443	0.792	0.836	0.576	0.697	0.793	3.248	0.719	0.624	0.454	0.584	
TR	0.371	1.628	0.479	0.441	0.792	0.835	0.576	0.697	0.792	3.15	0.724	0.627	0.464	0.588	
Bal	**0.383**	1.723	**0.488**	**0.454**	0.792	**0.838**	0.576	**0.699**	**0.797**	**3.373**	0.728	0.633	0.461	0.593	
PMLL	0.382	**1.727**	0.487	0.452	0.792	**0.838**	**0.577**	**0.699**	0.796	3.368	0.727	**0.636**	**0.471**	**0.597**	

Table 2: Different VLMs fine-tuning on COCO with joint captioning and classification tasks.

Task	Captioning										Classification				Model
Metric /Method	BLEU	CIDEr	R-L	SPICE	C-S	R-C -S	P-S++	R-P -S++	Polos	KLE	Head	Mid	Tail	Avg	
Softmax	0.532	3.935	0.662	0.610	0.792	**0.850**	0.579	**0.711**	0.795	4.840	0.862	0.817	0.762	0.809	**Qwen 3.5 -4B (Overall avg relative increase for PMLL 0.1)**
Focal	0.526	3.847	0.658	0.605	0.793	0.849	**0.580**	**0.711**	0.795	4.790	0.866	0.817	0.763	0.810	
CLA	0.514	3.755	0.653	0.593	**0.793**	**0.848**	**0.580**	0.709	0.790	4.515	**0.867**	0.809	0.734	0.796	
BS	0.480	3.377	0.628	0.575	0.791	0.845	0.578	0.706	0.780	4.314	0.856	0.811	0.763	0.805	
LinUpper	0.521	3.813	0.653	0.601	0.792	0.849	0.579	0.710	0.793	4.705	0.858	0.814	0.771	0.810	
TUL	**0.533**	3.944	**0.663**	**0.611**	0.792	**0.850**	0.579	**0.711**	0.795	4.862	0.862	0.817	0.759	0.808	
Gala	**0.533**	**3.952**	0.663	**0.611**	0.792	**0.850**	0.579	**0.711**	**0.796**	4.860	0.863	0.817	0.763	0.809	
TR	0.524	3.830	0.658	0.606	0.791	0.849	0.579	0.710	0.793	4.696	0.863	0.814	0.767	0.810	
Bal	**0.533**	3.941	0.663	0.610	0.792	0.849	**0.580**	**0.711**	0.795	**4.844**	0.864	0.817	0.759	0.808	
PMLL	**0.533**	3.943	**0.663**	**0.611**	**0.793**	**0.850**	**0.580**	**0.711**	**0.796**	**4.844**	0.861	**0.819**	**0.776**	**0.814**	
Softmax	0.512	3.681	0.646	0.598	**0.790**	**0.847**	**0.578**	**0.708**	0.788	4.674	0.866	0.821	0.772	0.815	**Intern VL -4B (Overall avg relative increase for PMLL 0.3)**
Focal	0.504	3.601	0.642	0.592	**0.790**	0.846	**0.578**	**0.708**	0.786	4.592	**0.869**	0.820	0.776	0.817	
CLA	0.504	3.619	0.645	0.588	**0.790**	0.845	0.577	0.707	0.783	4.452	0.863	0.811	0.742	0.799	
BS	0.477	3.348	0.625	0.573	0.789	0.843	0.576	0.705	0.776	4.323	0.846	0.809	0.769	0.804	
LinUpper	0.505	3.596	0.641	0.590	**0.790**	0.846	0.577	0.707	0.785	4.573	0.861	0.810	0.748	0.801	
TUL	0.514	3.721	0.648	0.600	**0.790**	**0.847**	0.577	**0.708**	**0.789**	4.684	0.864	0.819	0.758	0.808	
Gala	0.514	3.727	0.648	0.600	**0.790**	**0.847**	0.577	**0.708**	0.788	4.682	0.863	0.818	0.755	0.806	
TR	0.504	3.575	0.641	0.592	**0.790**	0.846	0.577	**0.708**	0.787	4.512	0.864	**0.826**	**0.785**	**0.821**	
Bal	**0.517**	**3.755**	**0.649**	**0.601**	**0.790**	**0.847**	0.577	**0.708**	**0.789**	**4.691**	0.863	0.818	0.756	0.807	
PMLL	0.514	3.721	**0.649**	**0.601**	**0.790**	**0.847**	**0.578**	**0.708**	**0.789**	**4.691**	0.862	0.821	0.780	0.817	

Extra Evaluation. VQA evaluation is conducted to assess the tail-concept and visual-detail learning for PMLL. As shown in Fig. 8, PMLL improves GT Containment and NLI P to G (Yao & Barbosa, 2024), indicating better label coverage. Although PMLL achieves lower NLI G to P scores, as shown in the examples, this is mainly because PMLL generates richer details that are not directly entailed by the ground-truth phrases. Such elaboration may not benefit short-answer VQA but supports PMLL's effectiveness in capturing tail concepts and visual details for captioning and classification.

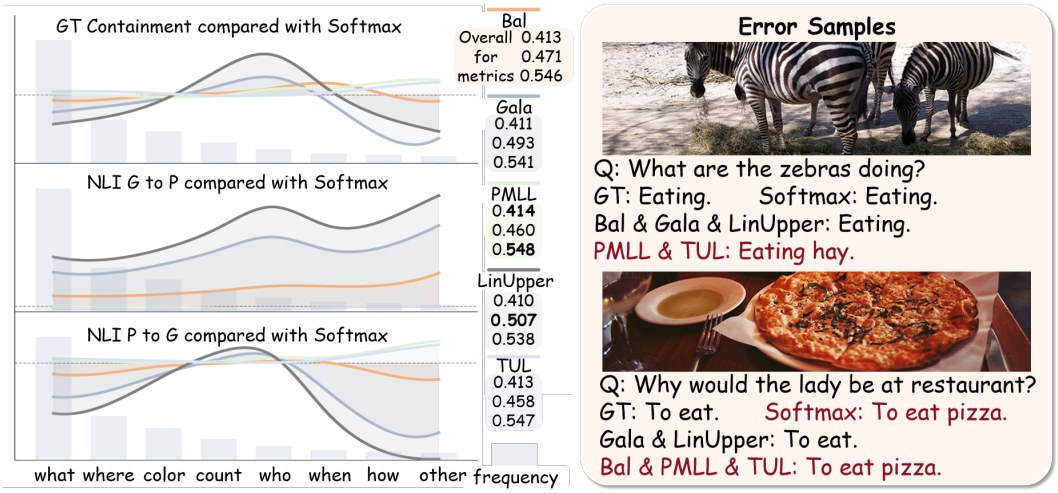


Figure 8: VQA results on VG dataset. GT Containment measures label inclusion. NLI G to P and P to G denote bidirectional entailment probabilities, respectively.

Table 3: Ablation study of PMLL. MLR, RWR, and ATO denote modality-decoupled long-tail re-weighting, rare word reinforcement, and adaptive task-aware optimization, respectively.

MLR	RWR	ATO	λ_V	λ_T	λ_J	BLEU	CIDEr	R-L	SPICE	Polos	KLE	Head	Mid	Tail	Avg
×	×	×	-	-	-	0.476	3.368	0.621	0.566	0.782	4.350	**0.861**	0.797	0.753	0.799
✓	×	×	0.25	0.25	0.5	0.476	3.375	0.621	0.566	0.782	4.353	0.860	**0.799**	0.762	0.802
✓	✓	×	0.25	0.25	0.5	**0.479**	3.405	**0.623**	**0.568**	**0.783**	4.374	0.859	0.798	0.761	0.802
✓	✓	✓	0.33	0.33	0.34	0.478	3.401	**0.623**	**0.568**	**0.783**	4.371	0.860	0.797	**0.765**	**0.803**
✓	✓	✓	0.1	0.1	0.8	**0.479**	3.404	**0.623**	**0.568**	**0.783**	**4.375**	0.860	0.795	0.764	0.802
✓	✓	✓	0.25	0.25	0.5	**0.479**	**3.406**	**0.623**	**0.568**	**0.783**	**4.375**	0.859	0.796	**0.765**	0.802

4.3 Ablation Studies

Ablation studies are conducted on COCO using Qwen2.5-32B as teacher and Qwen2.5-3b as student, to evaluate the effectiveness of each PMLL component. As shown in Tab. 3, MLR improves tail-class recognition by incorporating modality-specific long-tail information, while RWR further enhances captioning performance by emphasizing rare and class-related words. Incorporating ATO provides additional benefits for tail-class recognition, demonstrating the complementary roles of the three components in joint captioning and classification. We also investigate the sensitivity to different modality weights. PMLL maintains comparable performance across the evaluated configurations, indicating that it is relatively robust to moderate variations in λ_V, λ_T, and λ_J. We adopt $(0.25, 0.25, 0.5)$ as the default setting for all experiments, which achieves a balanced performance across the two tasks.

5 Conclusion.

We investigate modality-inconsistent long-tail distributions in realistic multimodal VLM learning and propose PMLL, a pre-decoupled multimodal-aware long-tail learning framework to address this problem. PMLL combines modality-decoupled long-tail re-weighting, rare word reinforcement, and adaptive task-aware optimization to improve rare-concept learning across captioning and classification tasks. A co-occurrence-aware tail mining protocol and KLE are further introduced to better reveal and evaluate long-tail effects for VLM learning. Extensive experiments demonstrate the effectiveness and robustness of PMLL in modality-inconsisten learning. Nevertheless, PMLL remains challenged by extreme tail concepts, ambiguous visual evidence, and strong concept co-occurrence. Its reliance on a predefined concept space and teacher-derived statistics may also limit open-vocabulary generalization. Future work will explore more fine-grained representation learning to better handle these challenging cases and extend PMLL to broader open-vocabulary settings.

REFERENCES

Jianing An, Luyang Jiang, Jie Luo, Wenjun Wu, and Lei Huang. Voqa: Visual-only question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9100–9109, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jiahao Chen, Bin Qin, Jiangmeng Li, Hao Chen, and Bing Su. Rethinking the Bias of Foundation Model Under Long-Tailed Distribution. In *International Conference on Machine Learning*, volume 267, pp. 8910–8930, 2025.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, November 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 105–115, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8577–8584, 2019.

Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, Yang Yang, and Heng Tao Shen. Generalizing vision-language models to novel domains: A comprehensive survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

Teli Ma, Shijie Geng, Mengmeng Wang, Sheng Xu, Hongsheng Li, Baochang Zhang, Peng Gao, and Yu Qiao. Unleashing the potential of vision-language models for long-tailed visual recognition. In *British Machine Vision Conference*, 2022.

Wenjun Miao, Mingda Li, Yanchao Hao, and Zheng Wei. Adaptive token refinement in long-tailed large vision-language models fine-tuning. In *International Conference on Machine Learning*, 2026.

Diego Ortego, Marlon Rodríguez, Mario Almagro, Kunal Dahiya, David Jiménez, and Juan C SanMiguel. Large language models meet extreme multi-label classification: scaling and multi-modal framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24630–24638, 2026.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12988–12997. IEEE, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for vision-and-language evaluation and training. *International Journal of Computer Vision*, 133(11):7647–7671, 2025.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Jiang-Xin Shi, Chi Zhang, Tong Wei, and Yu-Feng Li. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2663–2673, 2024.

Mingyang Song, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9434–9444. IEEE, 2025.

Daouda Sow, Herbert Woisetschläger, Saikiran Bulusu, Shiqiang Wang, Hans Arno Jacobsen, and Yingbin Liang. Dynamic loss-based sample reweighting for improved large language model pretraining. In *International Conference on Learning Representations*, volume 2025, pp. 89490–89520, 2025.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13559–13568, June 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, aug 2024b. ISSN 0920-5691.

Alexander Weers, Daniel Rueckert, and Martin J Menten. Weighting what matters: Boosting sample efficiency in medical report generation via token reweighting. In *Medical Imaging with Deep Learning*, 2026.

Yuexin Wu, I-Chan Huang, and Xiaolei Huang. Token imbalance adaptation for radiology report generation. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209, pp. 72–85, 22 Jun–24 Jun 2023.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 13066–13093, 2026.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Dragomir Anguelov. WOD-E2E: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3709–3718, 2026.

Peiran Yao and Denilson Barbosa. Accurate and nuanced open-qa evaluation through textual entailment. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2575–2587, 2024.

Mao Ye, Greg P Meyer, Zaiwei Zhang, Dennis Park, Siva Karthik Mustikovela, Yuning Chai, and Eric Wolff. Vlmine: Long-tail data mining with vision language models. In *Proceedings of the Winter Conference on Applications of Computer Vision Workshops*, pp. 1072–1082, 2025.

Fan Zhang, Wei Qin, Weijieying Ren, Lei Wang, Zetong Chen, and Richang Hong. Gradient-aware logit adjustment loss for long-tailed classifier. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3190–3194, 2024.

Jingyu Zhong, Ronghua Shang, Mengmeng Liu, Yuhao Xie, and Jie Feng. Noise correction and distribution fine-tuning for long-tailed partial multi-label learning. *Pattern Recognition*, 180: 114096, 2026.

A APPENDIX

A.1 ANALYSIS OF MODALITY-INCONSISTENT LONG-TAIL LEARNING

Data-level inconsistency. Conventional frequency-based reweighting assigns a concept-level weight according to its aggregated occurrence frequency as

$$w_c = f(n_c), \tag{8}$$

where n_c denotes the number of training samples containing concept c. This formulation implicitly assumes that the same distribution provides comparable learning evidence across modalities. However, the supervision available to a VLM depends not only on whether a concept occurs, but also on where it occurs and whether its visual and textual evidence is aligned. For this, let n_c^V, n_c^T, and n_c^J denote the numbers of samples in which concept c is observed in the visual modality, the textual modality, and post-feature-extraction joint multi-modality with aligned evidence, respectively. They satisfy

$$n_c^V \neq n_c^T \neq n_c^J \quad \text{in general.} \tag{9}$$

Specifically, two concepts with the same aggregated frequency may have substantially different modality-specific supervision. To illustrate this distinction, consider two hypothetical concepts c_a and c_b, each appearing in 100 training samples. Suppose their modality-specific occurrence counts for different modalities are

$$\begin{aligned}(n_{c_a}^J, n_{c_a}^V, n_{c_a}^T) &= (100, 150, 5),\\ (n_{c_b}^J, n_{c_b}^V, n_{c_b}^T) &= (100, 40, 70).\end{aligned} \tag{10}$$

Although $n_{c_a}^J = n_{c_b}^J = 100$, concept c_a has substantially fewer textual observations and aligned visual-textual instances. A weighting scheme based solely on n_c^J assigns identical weights to these two concepts, even though the evidence available for learning their cross-modal associations differs considerably. This example establishes that aggregated frequency alone is insufficient to characterize modality-specific concept scarcity. Moreover, occurrence does not necessarily imply independent supervision. A concept may repeatedly appear alongside dominant concepts, making its nominal frequency a potentially misleading indicator of independently available learning evidence. Thus, the underlying long-tail distribution can vary across visual recognition, textual generation, and cross-modal association, even when all three tasks are trained on the same dataset, resulting in a complex modality-inconsistent long-tail pattern.

Feature-level inconsistency. Differences in supervision quantity can further lead to differences in learned feature representations. To examine this relationship, consider the feature vectors $\{z_{i,c}^m\}_{i=1}^{n_c^m}$ associated with concept c in modality $m \in \{V, T, J\}$. For the following analysis, assume that these vectors are independent observations with population mean μ_c^m and finite covariance matrix Σ_c^m. Specifically, under the stated assumptions, the expected squared estimation error of the empirical concept prototype $\hat{\mu}_c^m$ is defined as

$$\mathbb{E}\left[\|\hat{\mu}_c^m - \mu_c^m\|_2^2\right] = \frac{\mathrm{tr}(\Sigma_c^m)}{n_c^m}, \quad \hat{\mu}_c^m = \frac{1}{n_c^m}\sum_{i=1}^{n_c^m} z_{i,c}^m. \tag{11}$$

It shows that the number of available observations and the dispersion of their features can affect representation reliability, suggesting that a large occurrence count does not guarantee a reliable representation when the corresponding features remain highly dispersed. Moreover, compact features estimated from very few observations may still have considerable statistical uncertainty.

For this reason, we exam feature compactness alongside occurrence frequency for modeling modality-inconsistent long-tail. For modality m, compactness can be expressed via feature extractor as

$$q_c^m = \left(\epsilon + \frac{1}{n_c^m}\sum_{i=1}^{n_c^m} \left\| z_{i,c}^m - \hat{\mu}_c^m \right\|_2^2\right)^{-1}. \tag{12}$$

Specifically, its within-concept dispersion provides an empirical indication of the covariance term of the expected squared estimation error. By doing this, frequency and compactness capture complementary evidence from occurrence quantity and the consistency of the resulting representations

Table 4: Results of VLMs fine-tuning on VG with joint captioning and classification tasks.

Task	Captioning										Classificaiton				Model
Metric /Method	BLEU	CIDEr	R-L	SPICE	C-S	R-C -S	P-S++	R-P -S++	Polos	KLE	Head	Mid	Tail	Avg	
Softmax	**0.421**	2.332	**0.549**	**0.515**	**0.791**	**0.836**	**0.580**	**0.702**	**0.797**	3.503	0.938	0.913	0.912	0.921	**Gemma4 to Qwen2.5 (Overall avg relative increase for PMLL 0.1)**
Focal	0.388	2.052	0.524	0.491	0.790	0.833	0.579	0.699	0.791	3.261	0.910	0.861	0.824	0.861	
CLA	0.402	2.137	0.533	0.489	0.789	0.832	0.579	0.699	0.791	3.143	**0.945**	0.908	0.899	0.916	
BS	0.373	1.894	0.507	0.459	0.784	0.825	0.575	0.692	0.777	2.906	0.928	0.901	0.906	0.911	
LinUpper	0.392	2.075	0.524	0.492	0.789	0.832	0.578	0.698	0.788	3.277	0.942	0.913	0.909	0.920	
TUL	**0.421**	2.331	**0.549**	0.515	**0.791**	**0.836**	**0.580**	**0.702**	0.797	3.508	0.937	0.913	0.911	0.920	
Gala	0.419	2.323	0.548	0.514	**0.791**	**0.836**	**0.580**	**0.702**	0.797	3.500	0.937	0.913	0.912	0.920	
TR	0.413	2.248	0.543	0.508	**0.791**	0.835	**0.580**	0.701	0.795	3.384	0.939	**0.919**	**0.918**	**0.925**	
Bal	**0.421**	2.362	0.543	0.514	**0.791**	**0.836**	**0.580**	**0.702**	0.797	3.526	0.929	0.903	0.904	0.912	
PMLL	**0.421**	**2.379**	**0.549**	**0.515**	**0.791**	**0.836**	**0.580**	**0.702**	**0.798**	**3.541**	0.932	0.910	0.917	0.920	

for concept learning. Importantly, compactness should not be interpreted as a sufficient measure of representation quality. A concept represented by only one observation has zero empirical within-concept dispersion, while representational collapse may also produce artificially compact features. Furthermore, compactness alone does not measure separation from competing concepts. It is therefore meaningful as a complementary signal considered jointly with occurrence statistics, rather than as an independent guarantee of reliable concept learning.

Additionally, the effect can be especially relevant to modality-inconsistent long-tail concepts. For example, a concept with abundant visual occurrences but limited textual or aligned evidence may contribute frequently to classification supervision while receiving relatively few concept-specific captioning updates. Its shared representation can therefore be shaped by task-dependent quantities and directions of learning signals. Conversely, concepts with substantial textual evidence but sparse visual grounding may receive strong language-related updates without correspondingly strong visual supervision. Since captioning also includes many tokens unrelated to the target concept, the volume of captioning supervision does not necessarily translate into an equivalent amount of concept-specific cross-modal learning.

Multi-task amplification for modality-inconsistent long-tail. Modality-inconsistent supervision is particularly consequential for VLMs since multi-tasks optimize different objectives while sharing model parameters. Specifically, consider their joint optimization as

$$\mathcal{L}(\theta) = \lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{a}}(\theta) + \lambda_{\mathrm{b}}\mathcal{L}_{\mathrm{b}}(\theta), \tag{13}$$

where θ denotes shared parameters. Let $g_{\mathrm{task}} = \nabla_\theta \mathcal{L}_{\mathrm{task}}$. For a gradient-descent step with learning rate η, a first-order expansion gives

$$\Delta\mathcal{L}_{\mathrm{a}} \approx -\eta\, g_{\mathrm{a}}^{\top}\,(\lambda_{\mathrm{a}} g_{\mathrm{a}} + \lambda_{\mathrm{b}} g_{\mathrm{b}}) \;= -\eta\lambda_{\mathrm{a}}\|g_{\mathrm{a}}\|_2^2 - \eta\lambda_{\mathrm{b}} g_{\mathrm{a}}^{\top} g_{\mathrm{b}}, \tag{14}$$

showing that progress on one task depends not only on its own gradient, but also on its interaction with the other task. In particular, if $\lambda_{\mathrm{a}} g_{\mathrm{a}}^{\top} g_{\mathrm{b}} < -\lambda_{\mathrm{a}}\|g_{\mathrm{b}}\|_2^2$, the combined first-order update increases the classification loss.

Overall, these observations identify three sources of modality-inconsistent long-tail learning, including unequal modality-specific occurrence counts, unequal feature-estimation reliability, and unequal or potentially conflicting task-specific optimization signals. Their combined effect cannot be fixed via aggregated concept frequency alone. This motivates our consideration of modality-specific occurrence and representation statistics, together with task-aware optimization, to address different forms of concept scarcity within a unified VLM.

A.2 Additional Experiments with Gemma-4 Teacher Model

To further validate the robustness of PMLL with different teacher models, we conduct additional experiments using Gemma-4-26B as the teacher model for concept extraction and long-tail statistics construction. The results are shown in Tab. 4. The results demonstrate that PMLL consistently improves the overall performance in caption and classification tasks with different teacher models, indicating that the proposed modality-aware long-tail modeling is not dependent on a specific teacher architecture.

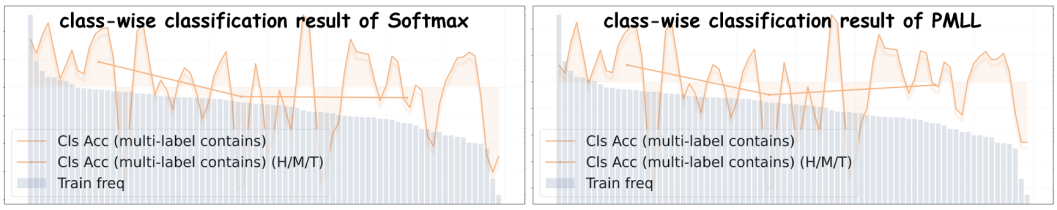


Figure 9: Class-wise classification and captioning performance across baselines after applying metric-agnostic co-occurrence-aware tail mining protocol.

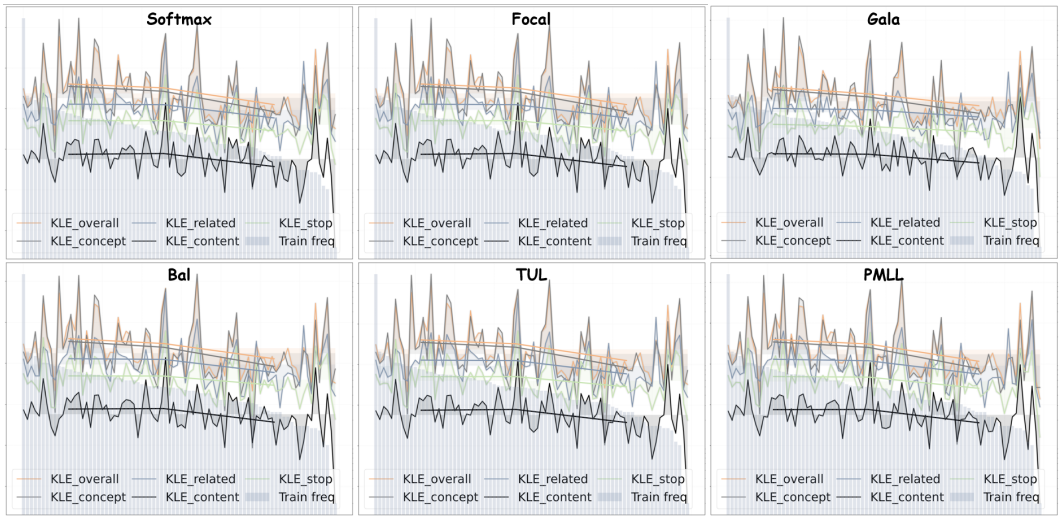


Figure 10: Class-wise KLE evaluation across baselines on COCO with Qwen model.

Moreover, Fig. 9 presents class-wise classification and captioning performance across baselines.It can be seen that even with the proposed metric-agnostic co-occurrence-aware tail mining protocol, multimodal learning still exhibits more complex long-tail patterns than conventional image classification, further highlighting the impact of modality-inconsistent long-tail on VLM learning from realistic multimodal datasets and underscore the importance of addressing this challenge.

A.3 Word-Type-Specific Component Analysis of KLE

In addition to evaluating the overall agreement between generated captions and ground-truth references, KLE incorporates word-type-specific consistency checks, as shown in Fig. 10. The results reveal substantial differences in long-tail distributions across word types, as well as varying performance of different methods on each word type. These findings further highlight the complexity of modality-inconsistent long-tail learning and suggest directions for more fine-grained multimodal long-tail learning.

A.4 Failure Case Analysis

Fig. 11 shows some failure cases of PMLL on COCO with Qwen model. Specifically, we present three examples in which the predicted labels differ substantially from the ground-truth labels. First, despite its larger model size, the teacher model does not always produce accurate predictions and may miss tail concepts that are uncommon or difficult to distinguish in images or text. In contrast, the student model fine-tuned with PMLL can recover some of these missing tail concepts, demonstrating the effectiveness of PMLL for multimodal model fine-tuning.

Second, some extremely rare or highly modality-inconsistent concepts remain difficult to recognize even after reweighting, particularly when they are occluded, truncated, or occupy only a small image region. This does not necessarily indicate a failure of PMLL, but showing that its ability to

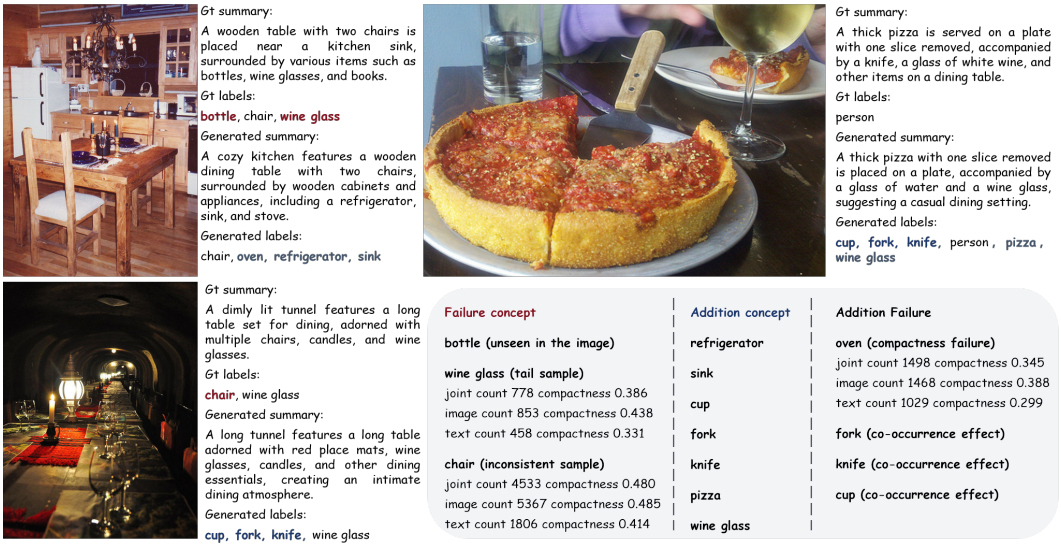


Figure 11: Analysis for some large label gap cases.

handle such challenging cases could be further improved through dedicated occlusion-aware training data or additional feature extraction mechanisms. In other words, the long-tail decoupling and tail-enhancement mechanisms in PMLL provide a robust solution for general modality-inconsistent long-tail scenarios but are not specifically designed for extreme samples or tasks.

Finally, PMLL does not introduce a dedicated module or method for addressing concept co-occurrence. Instead, we propose only a coarse co-occurrence-aware mining protocol for evaluation, which does not fully capture the underlying co-occurrence patterns. As a result, some strongly co-occurring concepts remain prone to misclassification.

For the captioning task, PMLL tends to generate more detailed descriptions or more comprehensive summaries. Overall, PMLL effectively alleviates modality-inconsistent long-tail learning and can even outperform the teacher model in concept generation in some cases. However, because it does not explicitly address anomalous or extreme samples, recognition errors may still occur in these challenging scenarios.

A.5 Implementation Details

Training Details. We fine-tune the Qwen2.5-VL-3B model using parameter-efficient LoRA adaptation. Specifically, LoRA modules are applied to the query, key, value, and output projection layers, i.e., q_{proj}, k_{proj}, v_{proj}, and o_{proj}. The LoRA rank and scaling factor are both set to 2. For optimization, we train the model for 3 epochs with a maximum sequence length of 4096 tokens. The learning rate is set to 5×10^{-5} with a warm-up ratio of 0.03. The per-device batch size is set to 4 with gradient accumulation over 4 steps, resulting in an effective batch size of 16. Gradient clipping is applied with a maximum norm of 0.5. All experiments are conducted with FP16 precision for efficient multimodal training. Images are resized online during data loading with the maximum edge length set to 224 pixels, while the original images remain unchanged. The dataloader uses 4 workers, and all experiments are performed with a fixed random seed of 42 to ensure reproducibility.

Loss Reweighting Implementation. For long-tail analysis, we categorize categories into head, medium, and tail groups according to their natural occurrence frequencies in each dataset. Specifically, all categories are sorted in descending order based on their training sample numbers. Following a 3/3/4 partition strategy, the top 30%, middle 30%, and remaining 40% categories are respectively assigned to the head, middle, and tail groups. Since all evaluated datasets naturally contain category imbalance, we do not manually construct additional long-tail splits or enforce predefined imbalance ratios. Instead, we directly preserve the original data distributions, which better reflects realistic multimodal long-tail scenarios where imbalance arises naturally.

Table 5: Results of VLMs fine-tuning on COCO with joint captioning and classification tasks under Random Seed as 0.

Task	Captioning										Classificaiton				Model
Metric /Method	BLEU	CIDEr	R-L	SPICE	C-S	R-C -S	P-S++	R-P -S++	Polos	KLE	Head	Mid	Tail	Avg	
Softmax	0.477	3.377	0.622	0.566	**0.793**	0.846	**0.580**	0.708	0.783	4.354	0.864	0.795	0.755	0.800	**Qwen2.5 -32B to 3B (Overall avg relative increase for PMLL 0.2)**
Focal	0.442	3.074	0.596	0.540	0.791	0.843	0.579	0.707	0.776	4.042	0.850	0.773	0.714	0.773	
CLA	0.457	3.168	0.609	0.553	0.792	0.844	**0.580**	0.706	0.777	4.118	**0.865**	0.792	0.738	0.793	
BS	0.433	2.924	0.591	0.535	0.791	0.842	0.578	0.704	0.768	3.924	0.847	0.791	0.753	0.793	
LinUpper	0.466	3.268	0.612	0.558	0.792	0.845	0.579	0.707	0.779	4.249	0.862	**0.803**	0.755	0.801	
TUL	0.477	3.379	0.622	0.566	**0.793**	0.846	**0.580**	**0.709**	0.783	4.350	0.862	0.796	0.752	0.799	
Gala	0.477	3.379	0.622	0.567	**0.793**	0.846	**0.580**	0.708	0.783	4.356	**0.865**	0.797	0.754	0.800	
TR	0.468	3.290	0.616	0.562	0.792	0.845	0.579	0.707	0.780	4.190	0.851	0.795	0.753	0.795	
Bal	0.478	3.384	**0.623**	**0.569**	**0.793**	**0.847**	**0.580**	0.708	0.783	4.354	0.864	0.800	0.753	0.801	
PMLL	**0.479**	**3.398**	**0.623**	0.568	**0.793**	**0.847**	**0.580**	**0.709**	**0.784**	**4.367**	0.861	0.797	**0.760**	**0.802**	

KLE Implementation. For the implementation of KLE, we empirically set the weights $\omega_t \in \{w_{cls}, w_{rel}, w_{con}, w_{stop}\}$ to 0.6, 0.3, 0.1, and 0, respectively. These values are assigned based on the relative importance of different part-of-speech categories defined in the preset vocabulary. The weighting scheme is heuristic and can be flexibly adjusted by modifying the part-of-speech categories or their corresponding weights according to specific application scenarios.

Computational cost. Taking the fine-tuning of Qwen2.5-3B on COCO using an NVIDIA RTX 3090 GPU as an example, standard softmax fine-tuning takes approximately 70 minutes, whereas both Balanced Softmax and PMLL require approximately 90 minutes. Computing rare word reinforcement and concept compactness takes less than 5 minutes in total, corresponding to less than 5.6% of the fine-tuning time. Thus, PMLL incurs modest additional computation while matching the fine-tuning time of Balanced Softmax. Moreover, PMLL only modifies loss weighting without changing the student architecture or introducing additional trainable parameters. Its primary additional computational cost arises from teacher inference for estimating the three modality-specific long-tailed distributions. Our procedure uses one image-conditioned pass to obtain the joint- and image-modality distributions and one text-only pass to obtain the text-modality distribution. Relative to a single-pass distribution estimation procedure, this entails approximately twice the number of teacher inference passes, although the actual runtime depends on the teacher model and inference configuration. This preprocessing cost could potentially be reduced through optimized API-based inference or a prompting strategy that produces all three distributions in a single pass. With a sufficiently comprehensive synonym lexicon, lexical matching could also provide an approximation to the text-modality distribution, reducing the need for teacher inference. Overall, the measured results demonstrate modest computational overhead during fine-tuning, with no additional student parameters.

A.6 Robustness Analysis under Different Random Seeds

To further validate the effectiveness and robustness of PMLL in addressing modality-inconsistent long-tail learning, we conduct additional experiments with a different random seed on the COCO dataset. Specifically, we use Qwen2.5-32B as the teacher model and Qwen2.5-3B as the student model, while keeping all other experimental settings unchanged. As shown in Table 5, PMLL consistently achieves the best overall performance across both captioning and classification tasks under different random seeds, demonstrating its robustness against training randomness. Importantly, although the improvements of PMLL over the best-performing baseline on individual metrics are relatively modest, jointly optimizing caption generation and classification generation remains challenging due to the inherent trade-off between heterogeneous task objectives. Therefore, maintaining consistently competitive performance across both tasks is more meaningful than optimizing a single metric independently. The stable improvements achieved by PMLL under different random seeds further verify that the proposed modality-aware long-tail modeling effectively enhances the overall capability of VLM fine-tuning.

A.7 AI Usage Statement

LLMs were used to assist with translation, writing polishing, and grammar checking.