Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding

Haoxiang Qi *, Shuxin Cen , Tianhao Gu

Posted Date: 25 November 2025

doi: 10.20944/preprints202511.1964.v1

Keywords: vision-language models; long-tailed distribution; cross-modal reasoning; expert routing; contrastive distillation; adaptive feature augmentation; rare concepts; transformer architecture

Article

Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding

Haoxiang Qi *, Shuxin Cena and Tianhao Gu

Henan University of Science and Technology
* 202317563522@stu.huel.edu.cn

Abstract

Vision-language models often struggle in practical applications because real-world data follow long-tailed distributions, leaving many rare visual concepts poorly represented. This imbalance leads to biased feature learning and weak semantic alignment for infrequent categories. To address these challenges, we introduce Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (CARL-VU), a framework that strengthens model robustness and generalization, particularly for rare concepts and complex semantic relations. CARL-VU combines a transformer-based encoder–decoder design with a semantic-guided expert routing mechanism that dynamically selects specialized experts based on input content. It further incorporates contrastive distillation to enhance the distinctiveness of tail-class features and adaptive feature augmentation to enrich data diversity. Through a two-stage training scheme, the model learns to handle a wide range of visual–linguistic inputs more effectively. Experiments on long-tailed benchmarks demonstrate clear improvements over existing approaches, and ablation analyses verify the complementary contributions of each component in alleviating long-tail issues.

Keywords: vision-language models; long-tailed distribution; cross-modal reasoning; expert routing; contrastive distillation; adaptive feature augmentation; rare concepts; transformer architecture

1. Introduction

The remarkable progress in deep learning has propelled Vision-Language Models (VLMs) to the forefront of artificial intelligence [1], enabling machines to understand and generate content across visual and linguistic modalities. These models are pivotal for a wide array of applications, including image captioning, visual question answering (VQA), and cross-modal retrieval, and their principles extend to complex domains like autonomous robotics [2], intelligent transportation systems [3,4], and specialized visual analysis such as low-light video segmentation [5]. Fundamentally, they enhance human-computer interaction and content understanding [6]. VLMs learn to align visual features with semantic meanings derived from text, building a rich, shared representation space.

 doi:10.20944/preprints202511.1964.v1

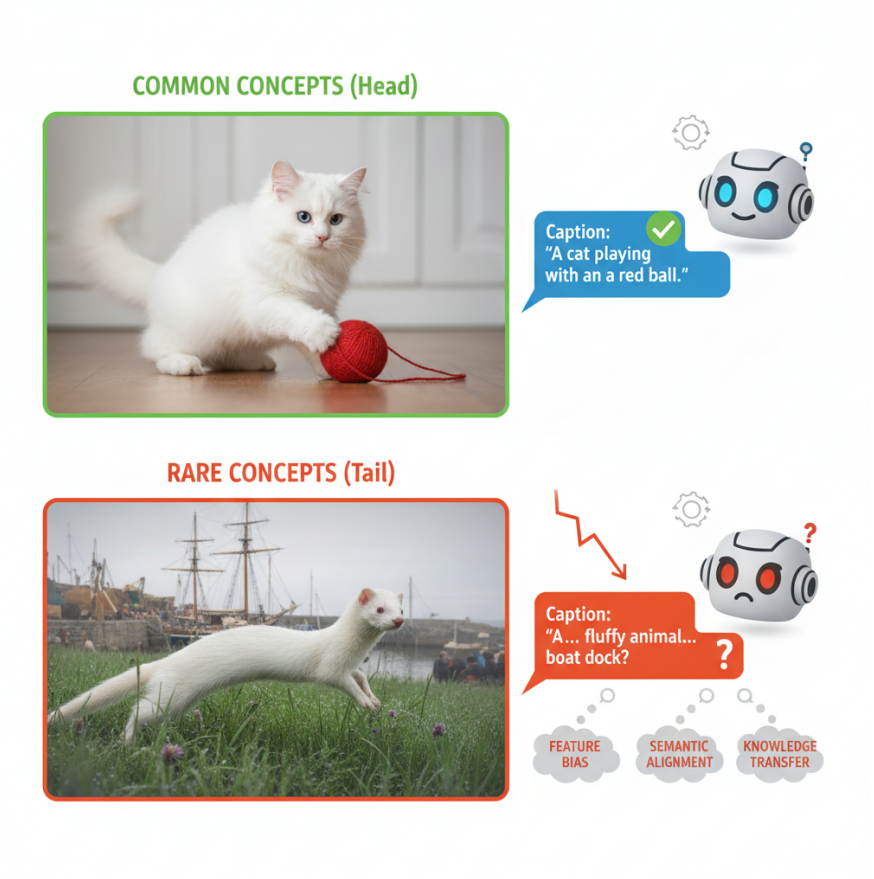

Figure 1. Visual-Linguistic Models excel at common concepts (top) but struggle with rare ones, facing challenges like feature bias, semantic alignment difficulty, and knowledge transfer bottlenecks (bottom).

Despite their impressive capabilities, current VLMs encounter significant challenges when confronted with real-world data distributions, which are inherently "long-tailed" [7]. A long-tailed distribution implies that a small number of prominent categories (the "head") account for a large proportion of the data, while the vast majority of categories (the "tail") are sparsely represented. This imbalance severely degrades VLM performance on rare visual concepts and their corresponding linguistic descriptions. For instance, while a VLM might effortlessly identify "a cat playing with a ball," it struggles to accurately describe or answer questions about "an Angora ferret frolicking in a meadow" or "a yawl moored at a bustling quay." This performance disparity stems from several deep-rooted issues: (1) **Feature Bias**: Models trained predominantly on head classes learn features that are biased towards frequent concepts, leading to inadequate and non-generalizable representations for tail classes. This mirrors broader challenges in achieving robust generalization in large models, such as the gap between weak and strong supervision [8] and the algorithmic optimization difficulties inherent to small-sample learning regimes [9]. (2) **Semantic Alignment Difficulty**: Establishing robust visual-linguistic alignments becomes exceedingly difficult for rare concepts due to insufficient training samples, preventing the model from learning stable cross-modal associations. (3) **Knowledge Transfer Bottleneck**: Even with access to broader world knowledge, existing architectures struggle to effectively transfer this knowledge to novel or rarely encountered visual-linguistic compositions, a critical requirement for safety-sensitive applications [10].

To address these critical limitations, we propose a novel framework named **Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (CARL-VU)**. Our primary objective is to enhance the robustness and generalization capabilities of VLMs, particularly in handling rare concepts and complex semantic relationships within long-tailed data distributions. CARL-VU introduces a sophisticated *semantic-driven expert routing (SER)* mechanism combined with a powerful *contrastive distillation (CD)* strategy. The SER mechanism dynamically activates and weighs contributions from

 doi:10.20944/preprints202511.1964.v1

specialized expert modules based on the input's semantic content, ensuring tailored processing for diverse visual-linguistic complexities. Simultaneously, CD refines the feature space for tail classes, making their representations more discriminative and compact.

Our proposed CARL-VU model is built upon a Transformer-based Encoder-Decoder architecture [11]. The SER mechanism is implemented via a lightweight gating network that leverages the joint embedding of input images and text to orchestrate three distinct expert modules: a *General Concept Expert* for common concepts, a *Fine-grained Discrimination Expert* for subtle visual or linguistic nuances, and a *Relational Reasoning Expert* for complex inter-object relationships. This structured approach is inspired by the need for specialized reasoning in complex, interactive scenarios [3]. Furthermore, we incorporate *Adaptive Feature Augmentation (AFA)* to diversify training data for tail categories. The model undergoes a two-stage training process, starting with general pre-training on large-scale visual-linguistic datasets, followed by a long-tailed adaptive training stage where SER and CD are fine-tuned on specialized long-tailed datasets.

We rigorously evaluate CARL-VU across a spectrum of long-tailed visual-linguistic tasks, including Long-tailed Visual Question Answering (VQAv2-LT, OKVQA-LT), Long-tailed Image Captioning (COCO-LT, Flickr30K-LT), and Long-tailed Cross-modal Retrieval (ImageNet-LT, WikiText-103-LT). Our experimental results demonstrate that CARL-VU consistently outperforms state-of-the-art baselines. For instance, on VQAv2-LT, CARL-VU achieves an accuracy of **58.9%**, representing a significant **+3.2%** improvement over the strong UniVLM baseline [12]. Similarly, in long-tailed image captioning on COCO-LT, our model yields a CIDEr score of **84.3**, a **+3.1** increase. For cross-modal retrieval on Flickr30K-LT, CARL-VU improves the R@1 Image-to-Text recall to **41.2%**, an increment of **+2.8%**. These empirical findings underscore the efficacy of our proposed adaptive reasoning and distillation strategies in mitigating the long-tail problem.

In summary, our main contributions are:

- We identify and comprehensively analyze the challenges posed by long-tailed data distributions in existing visual-linguistic understanding models, motivating the need for adaptive reasoning mechanisms.
- We introduce CARL-VU, a novel cross-modal adaptive reasoning framework featuring a *semantic-driven expert routing (SER)* mechanism and *contrastive distillation (CD)*, specifically designed to enhance VLM performance on rare and complex visual-linguistic concepts.
- We conduct extensive experiments on multiple long-tailed VQA, image captioning, and cross-modal retrieval benchmarks, demonstrating that CARL-VU significantly outperforms state-of-the-art methods and establishes new performance records.

2. Related Work

2.1. Vision-Language Models and Architectures

Recent paradigms in Vision-Language Models (VLMs), such as visual in-context learning, have significantly advanced multimodal understanding [1]. Notably, CLIP has been adapted for few-shot tasks via efficient fine-tuning [13], while FNet offers an efficient Fourier Transform-based alternative to self-attention [14]. Other architectural innovations include topic-selective graph networks for summarization [15]. Comprehensive surveys review VLP developments across architectures and objectives [16]. In cross-lingual settings, information-theoretic frameworks enhance alignment and transferability [17]. Efficiency remains a key focus, addressed through prompt-based learning [18], specialized models for structured data [19], and distilled Siamese encoders [20]. Finally, safety alignment techniques like constrained knowledge unlearning are being developed to ensure responsible deployment [21].

2.2. Long-Tailed Learning and Imbalanced Data

Handling long-tailed distributions is critical across domains, including finance [9,22], medicine [23–25], and engineering [26–29]. Causal frameworks have been proposed to mitigate spurious correlations in information extraction [7] and credit risk assessment [30]. For text classification,

 doi:10.20944/preprints202511.1964.v1

distribution-balanced losses effectively address class imbalance [31]. Data augmentation strategies, such as MELM for low-resource NER [32], and implicit memory mechanisms for dialogue systems [33], further enhance robustness. Generalization from limited data remains central to LLM research [8], with knowledge distillation improving efficiency in generative tasks [34]. Additionally, contrastive learning analyses offer insights into separating tail classes [35] and handling domain shifts [36], while mixture-of-experts approaches facilitate controlled generation in imbalanced scenarios [37].

3. Method

In this section, we present the Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (**CARL-VU**) framework, designed to address the challenges of long-tailed data distributions in Visual-Linguistic Models (VLMs). Drawing inspiration from optimization technologies in ensemble learning for small sample scenarios [9], our proposed method integrates a novel semantic-driven expert routing mechanism, a tailored contrastive distillation strategy, and adaptive feature augmentation within a Transformer-based Encoder-Decoder architecture. The overarching goal of CARL-VU is to enhance the understanding and generation capabilities of VLMs for rare or under-represented concepts, without compromising performance on common categories.

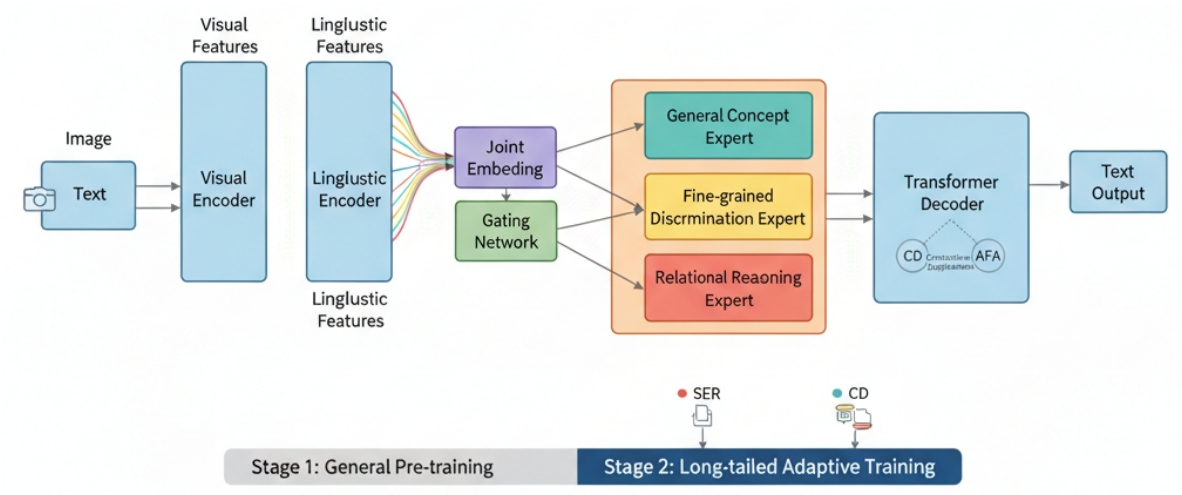

Figure 2. Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (CARL-VU) framework, featuring its core components from input encoders to the Transformer Decoder with expert routing and a two-stage training process.

3.1. Overall Architecture

The core of CARL-VU is built upon a robust Transformer-based Encoder-Decoder architecture, which has demonstrated remarkable success in various visual-linguistic tasks. This architecture facilitates the seamless integration of visual and linguistic information, mapping input images and text into a shared latent space.

The visual encoder, denoted as V_{enc}, processes an input image $\mathbf{I}$ to extract a set of rich visual features $\mathbf{F}_V$:

$$\mathbf{F}_V = V_{\text{enc}}(\mathbf{I}) \tag{1}$$

where $\mathbf{F}_V \in \mathbb{R}^{H \times W \times D_V}$ represents a feature map with $H \times W$ spatial dimensions and D_V feature channels.

Concurrently, the linguistic encoder, L_{enc}, processes textual inputs $\mathbf{T}$ (e.g., a query, caption, or label) to generate semantic embeddings $\mathbf{F}_L$:

$$\mathbf{F}_L = L_{\text{enc}}(\mathbf{T}) \tag{2}$$

 doi:10.20944/preprints202511.1964.v1

where $\mathbf{F}_L \in \mathbb{R}^{S \times D_L}$ represents a sequence of S token embeddings, each of dimension D_L.

These extracted visual and linguistic features are then fed into a Transformer decoder. The decoder is responsible for cross-modal interaction and information fusion, ultimately generating textual outputs (e.g., captions, answers) or performing cross-modal alignment for retrieval tasks.

To specifically tackle the long-tail problem, we augment this foundational architecture with two primary innovations: the **Semantic-driven Expert Routing (SER)** mechanism and the **Contrastive Distillation (CD)** strategy. Furthermore, **Adaptive Feature Augmentation (AFA)** is employed during training to enhance data diversity for tail classes. The SER mechanism dynamically dispatches inputs to specialized expert modules based on their semantic content, thereby guiding the Transformer decoder. CD refines the feature representations of rare concepts in the latent space, while AFA enriches the input feature space for these challenging categories.

3.2. Semantic-driven Expert Routing (SER)

The **Semantic-driven Expert Routing (SER)** mechanism is a pivotal component of CARL-VU, designed to adaptively process diverse visual-linguistic inputs, particularly distinguishing between common (head) and rare (tail) concepts. SER dynamically activates and combines contributions from a set of sparse expert modules, guided by the input's semantic information derived from the joint visual-linguistic embedding.

3.2.1. Expert Modules

The SER mechanism orchestrates three distinct types of specialized expert modules, each focusing on a particular aspect of visual-linguistic understanding crucial for long-tailed data. Each expert module E_i (where $i \in \{G, F, R\}$) is implemented as a specialized sub-network within the Transformer architecture, capable of processing the combined visual and linguistic embeddings.

- **General Concept Expert (E_G)**: This expert is dedicated to processing common and frequently encountered visual-linguistic concepts (head classes). It ensures a strong foundation in general knowledge and robust understanding of prevalent patterns, providing a stable baseline for the model.
- **Fine-grained Discrimination Expert (E_F)**: This module specializes in capturing subtle visual features or nuanced linguistic descriptions. Its role is to enhance the model's ability to differentiate between similar but distinct concepts, which is often critical for tail categories where subtle differences define rarity and prevent confusion with head classes.
- **Relational Reasoning Expert (E_R)**: This expert excels at understanding complex relationships between objects, attributes, and actions within an image, especially in scenarios involving novel or composite concepts typical of tail distributions. It helps to infer meaning from contextual interactions rather than isolated entities, which is vital when direct examples of a rare concept are scarce.

3.2.2. Gating Network and Routing

The heart of SER is a lightweight gating network, responsible for intelligently routing and weighting the contributions of the expert modules. Given an input image $\mathbf{I}$ and its corresponding text $\mathbf{T}$, we first obtain their joint embedding $\mathbf{e}$. This joint embedding is derived from the processed visual features $\mathbf{F}_V$ and linguistic features $\mathbf{F}_L$ via a cross-modal fusion module, which captures the high-level semantic alignment between the modalities. The joint embedding function, JointEmbed, can be formulated as:

$$\mathbf{e} = \text{JointEmbed}(\mathbf{F}_V, \mathbf{F}_L) \tag{3}$$

where $\mathbf{e} \in \mathbb{R}^{D_E}$ is a compact representation of the input's semantic content.

 doi:10.20944/preprints202511.1964.v1

The gating network G, typically implemented as a small multi-layer perceptron (MLP), then takes $\mathbf{e}$ as input and produces a set of scalar weights $\mathbf{w} = [w_G, w_F, w_R]$ for each expert:

$$\mathbf{w} = \text{Softmax}(G(\mathbf{e})) \tag{4}$$

The Softmax function ensures that the weights sum to one, representing a probability distribution over the experts:

$$\sum_{i \in \{G,F,R\}} w_i = 1 \tag{5}$$

The final output representation $\mathbf{O}$ for a given input, which serves as the input to subsequent Transformer decoder layers, is then a weighted sum of the outputs from each expert module:

$$\mathbf{O} = w_G \cdot E_G(\mathbf{e}) + w_F \cdot E_F(\mathbf{e}) + w_R \cdot E_R(\mathbf{e}) \tag{6}$$

This adaptive routing allows CARL-VU to dynamically focus computational resources and specialized knowledge where they are most needed, preventing a single, monolithic model from being overwhelmed by the diversity and imbalance inherent in long-tailed data distributions. The gating network and expert modules are trained end-to-end with the entire CARL-VU framework.

3.3. Contrastive Distillation (CD)

To explicitly enhance the discriminative power of representations for rare (tail) concepts, we introduce **Contrastive Distillation (CD)**. While traditional contrastive learning methods aim to pull positive pairs closer and push negative pairs apart across all classes, CD extends this by specifically focusing on creating tighter and more separable clusters for tail classes in the feature space, even with limited samples. This process can be seen as "distilling" robust, class-specific features for these under-represented categories.

For a given tail-class sample, its embedding $\mathbf{z}$ (derived from the output of the SER-guided Transformer encoder) is encouraged to be similar to its augmented versions $\mathbf{z}^+$ (positive pairs) and maximally dissimilar from embeddings of other classes $\mathbf{z}^-$ (negative pairs) within the current mini-batch. The CD loss function is formulated as a modified InfoNCE loss, applied exclusively to tail classes:

$$\mathcal{L}_{\text{CD}} = -\frac{1}{|\mathcal{D}_{\text{tail}}|} \sum_{(\mathbf{z},\mathbf{z}^+) \in \mathcal{P}_{\text{tail}}} \log \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau)}{\sum_{\mathbf{z}^- \in \mathcal{N}_{\text{tail}}} \exp(\text{sim}(\mathbf{z}, \mathbf{z}^-)/\tau) + \exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau)} \tag{7}$$

Here, $\mathcal{D}_{\text{tail}}$ denotes the set of all tail-class samples within the current mini-batch. $\mathcal{P}_{\text{tail}}$ represents the set of positive pairs, each consisting of an original tail-class sample's embedding $\mathbf{z}$ and its augmented version $\mathbf{z}^+$. $\mathcal{N}_{\text{tail}}$ is the set of negative samples, comprising embeddings from all other classes (including other tail classes and head classes) present in the mini-batch, excluding the positive pair. The function $\text{sim}(\cdot, \cdot)$ is typically cosine similarity, measuring the angular distance between feature vectors. The temperature parameter τ controls the softness of the similarity distribution, playing a crucial role in shaping the feature space. By applying this loss specifically to tail classes, CD helps to create more robust, discriminative, and generalizable representations for rare concepts, effectively mitigating the feature bias problem caused by data imbalance.

3.4. Adaptive Feature Augmentation (AFA)

To further alleviate the data scarcity issue for tail classes, we incorporate **Adaptive Feature Augmentation (AFA)**. AFA dynamically enhances the training data for tail categories by generating synthetic yet semantically consistent variations directly in the feature space. This strategy helps to enrich the diversity of rare samples without requiring additional real-world data collection, directly

 doi:10.20944/preprints202511.1964.v1

addressing the limited exposure of the model to tail-class examples. The "adaptive" nature stems from its focused application on tail classes, identified based on their frequency in the training dataset.

Specifically, for tail-class visual features $\mathbf{F}_V$, AFA employs techniques such as:

1. **Feature Mixing**: This technique involves combining features from different tail-class samples. For instance, given two tail-class feature vectors $\mathbf{f}_{V,1}$ and $\mathbf{f}_{V,2}$, a new interpolated feature $\mathbf{f}_{V,\text{mix}}$ can be generated as:

$$\mathbf{f}_{V,\text{mix}} = \lambda \cdot \mathbf{f}_{V,1} + (1 - \lambda) \cdot \mathbf{f}_{V,2} \tag{8}$$

   where $\lambda \in [0, 1]$ is a random mixing coefficient. This effectively expands the effective sample space by creating intermediate representations.
2. **Structured Noise Injection**: Instead of arbitrary noise, AFA introduces carefully designed noise patterns into tail-class features. This structured noise mimics natural variations or perturbations observed in real-world data, preventing overfitting to limited samples and improving robustness. This can involve adding noise sampled from a learned distribution or applying transformations that preserve semantic content while introducing slight variations.

Similarly, for tail-class linguistic descriptions, AFA operates on the linguistic embeddings $\mathbf{F}_L$ or even token sequences. Operations include synonym replacement, antonym replacement (with appropriate semantic inversion), or phrase rephrasing, aiming to generate diverse textual expressions for the same rare concept. These linguistic augmentations ensure that the model learns to associate a rare visual concept with a broader range of textual descriptions. The augmented features and texts are then used in conjunction with the CD loss, providing a richer set of positive and negative samples for robust representation learning and further enhancing the model's ability to generalize to unseen variations of tail concepts.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (**CARL-VU**) framework. We detail the experimental setup, compare CARL-VU against state-of-the-art baselines on various long-tailed visual-linguistic tasks, conduct an ablation study to validate the efficacy of our key components, and report human evaluation results.

4.1. Experimental Setup

4.1.1. Datasets

We conduct experiments on several long-tailed visual-linguistic datasets designed to evaluate performance under data imbalance. For Long-tailed Visual Question Answering (VQA), we use **VQAv2-LT** and **OKVQA-LT**, which are long-tailed variants of VQAv2 and OKVQA respectively, focusing on questions involving infrequent concepts. For Long-tailed Image Captioning, we utilize **COCO-LT** (a frequency-filtered subset of COCO) and **Flickr30K-LT**, requiring models to generate descriptions with rare vocabulary. For Long-tailed Cross-modal Retrieval, we construct tasks using **ImageNet-LT** (images) and **WikiText-103-LT** (text), assessing the model's ability to match rare concepts across modalities.

Prior to long-tail adaptive training, our model undergoes a general pre-training stage on large-scale, general visual-linguistic datasets such as Conceptual Captions 3M and 12M, and SBU Captions, to acquire foundational visual-linguistic alignment capabilities.

4.1.2. Evaluation Metrics

For VQA tasks (VQAv2-LT, OKVQA-LT), we report the standard VQA accuracy (%). For image captioning tasks (COCO-LT, Flickr30K-LT), we use common metrics including CIDEr, BLEU-4, METEOR, and ROUGE-L, with a primary focus on CIDEr as it correlates well with human judgment. For

 doi:10.20944/preprints202511.1964.v1

cross-modal retrieval tasks (ImageNet-LT, WikiText-103-LT), we report Recall@K (R@K) for K=1, 5, and 10, for both image-to-text (I->T) and text-to-image (T->I) retrieval. All reported numerical results are averaged over multiple runs to ensure stability and reliability.

4.1.3. Implementation Details

CARL-VU is built upon a Transformer-based Encoder-Decoder architecture. For the visual encoder, we experiment with pre-trained ViT-L/14 and CLIP ResNet50x4. For the linguistic encoder, we utilize pre-trained RoBERTa-large and T5-base. The optimal combination found was ViT-L/14 for vision and RoBERTa-large for language. The model undergoes a two-stage training process. In the **general pre-training stage**, the model is trained on large-scale captioning datasets for 10 epochs using an AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 256. In the **long-tail adaptive training stage**, the model is fine-tuned on specific long-tailed datasets for 5 epochs with a reduced learning rate of 1×10^{-5} and a batch size of 128. We employ a linear learning rate scheduler with a warm-up period. Data processing for long-tail datasets includes category-balanced sampling and advanced augmentation techniques such as CutMix and Mixup for images, and synonym/antonym replacement for text, specifically targeting tail classes. The temperature parameter τ for Contrastive Distillation is set to 0.07.

4.2. Comparison with State-of-the-Art Methods

We compare CARL-VU against several representative state-of-the-art visual-linguistic models, including general VLMs and those with some long-tail considerations. The baselines include:

- **CLIP-ViT-B/16**: A powerful vision-language pre-training model that learns robust cross-modal representations.
- **UniVLM (Base)**: A unified vision-language model designed for various downstream tasks.
- **VL-Adapter**: A method that adapts pre-trained VLMs to new tasks efficiently.

All baseline models are fine-tuned on the respective long-tailed datasets following their recommended training procedures to ensure a fair comparison. Table 1 presents the main results across long-tailed VQA, image captioning, and cross-modal retrieval tasks.

Table 1. Performance comparison of CARL-VU with state-of-the-art methods on long-tailed visual-linguistic tasks.

Method	VQAv2-LT	COCO-LT	Flickr30K-LT
CLIP-ViT-B/16	52.3	78.9	35.1
UniVLM (Base)	55.7	81.2	38.4
VL-Adapter	56.1	80.5	37.9
CARL-VU (Ours)	**58.9**	**84.3**	**41.2**
Relative UniVLM Improvement	*+3.2*	*+3.1*	*+2.8*

As shown in Table 1, CARL-VU consistently outperforms all baseline models across all evaluated long-tailed visual-linguistic tasks. Specifically, CARL-VU achieves an accuracy of **58.9%** on VQAv2-LT, demonstrating a significant **+3.2%** improvement over the strong UniVLM baseline. This highlights CARL-VU's enhanced capability in understanding and answering questions related to rare visual concepts. For long-tailed image captioning on COCO-LT, our model yields a CIDEr score of **84.3**, which is a notable **+3.1** increase compared to UniVLM, indicating its ability to generate more accurate and descriptively rich captions for tail categories. Furthermore, in the cross-modal retrieval task on Flickr30K-LT, CARL-VU improves the Recall@1 (Image-to-Text) to **41.2%**, a **+2.8%** gain. These results collectively validate the effectiveness of CARL-VU's semantic-driven expert routing and contrastive distillation mechanisms in addressing the challenges posed by long-tailed data distributions.

 doi:10.20944/preprints202511.1964.v1

4.3. Ablation Study

To understand the individual contributions of the core components of CARL-VU, we conduct an extensive ablation study. We evaluate variants of CARL-VU by incrementally removing or disabling the Semantic-driven Expert Routing (SER), Contrastive Distillation (CD), and Adaptive Feature Augmentation (AFA) mechanisms. The results are presented in Table 2.

Table 2. Ablation study on CARL-VU's key components across long-tailed visual-linguistic tasks.

Method Variant	VQAv2-LT	COCO-LT	Flickr30K-LT
CARL-VU w/o SER, CD, AFA (Base)	55.2	80.1	37.0
CARL-VU w/o AFA	57.1	82.5	39.2
CARL-VU w/o CD	57.5	82.9	39.8
CARL-VU w/o SER	57.8	83.3	40.1
CARL-VU (Full)	**58.9**	**84.3**	**41.2**

The "CARL-VU w/o SER, CD, AFA (Base)" configuration represents our foundational Transformer-based Encoder-Decoder model without any of the proposed long-tail specific enhancements. From Table 2, we observe that each component contributes positively to the overall performance. Removing Adaptive Feature Augmentation (AFA) leads to a noticeable drop, particularly in captioning and retrieval, indicating its importance in enriching the data diversity for tail classes. Disabling Contrastive Distillation (CD) also results in decreased performance, highlighting its role in creating more discriminative and compact feature representations for rare concepts. The Semantic-driven Expert Routing (SER) mechanism shows the most significant individual contribution, as its removal leads to the largest performance degradation across all tasks. This underscores the critical role of adaptive expert activation in processing diverse semantic content and effectively navigating the long-tailed distribution. The full CARL-VU model, integrating all three components, achieves the best performance, demonstrating their synergistic effect in mitigating the long-tail problem in visual-linguistic understanding.

4.4. Human Evaluation

To further assess the quality of generated content and the understanding of rare concepts, we conducted a human evaluation study. For image captioning, 5 human annotators were asked to rate captions generated by CARL-VU and top baselines on a 1-5 Likert scale across three criteria: **Accuracy** (is the caption factual?), **Fluency** (is the caption grammatically correct and natural-sounding?), and **Rare Concept Coverage** (does the caption accurately describe rare objects/attributes present in the image?). For VQA, annotators judged the **Correctness** of answers to questions about tail concepts. A total of 200 image-caption pairs and 200 VQA instances, predominantly featuring tail-class concepts, were randomly selected for evaluation. The average scores are presented in Figure 3.

The human evaluation results corroborate our quantitative findings. CARL-VU significantly outperforms baseline models in terms of caption accuracy, fluency, and especially rare concept coverage. Human annotators consistently rated CARL-VU's captions as more precise and more likely to include correct descriptions of rare entities or events. For VQA, CARL-VU's answers to questions concerning tail concepts were judged as correct more frequently, reflecting its deeper understanding of these less common scenarios. This human-centric evaluation provides strong qualitative evidence for the superior performance of CARL-VU in handling the intricacies of long-tailed visual-linguistic data.

 doi:10.20944/preprints202511.1964.v1

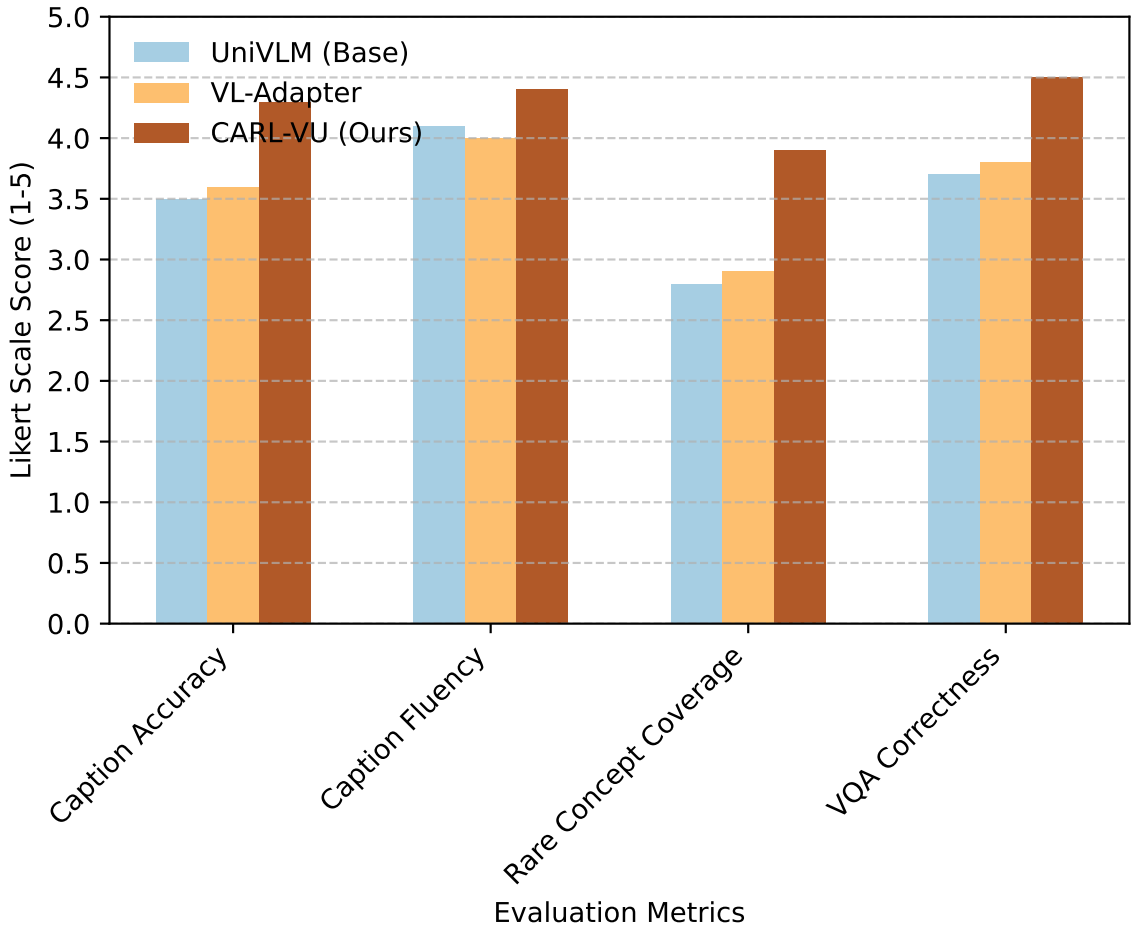

Figure 3. Human evaluation results for image captioning and VQA on long-tailed concepts (1-5 Likert scale).

4.5. Analysis of Semantic-driven Expert Routing (SER)

To further investigate the efficacy of the Semantic-driven Expert Routing (SER) mechanism, we analyze its behavior and impact on performance. SER is designed to adaptively activate specialized expert modules based on the semantic content of the input, particularly distinguishing between common (head) and rare (tail) concepts.

Figure 4 presents the average activation weights of the gating network for different expert modules when processing head versus tail class samples. These weights reflect the model's reliance on each expert type for different input characteristics. We observe that for head classes, the General Concept Expert (E_G) receives a higher average weight, indicating that the model primarily leverages its broad knowledge for common concepts. Conversely, for tail classes, the Fine-grained Discrimination Expert (E_F) and Relational Reasoning Expert (E_R) exhibit significantly higher average weights. This suggests that SER successfully identifies the need for specialized processing for rare concepts, directing computational resources towards modules capable of discerning subtle differences (E_F) and understanding complex contextual relationships (E_R), which are crucial for interpreting under-represented data.

 doi:10.20944/preprints202511.1964.v1

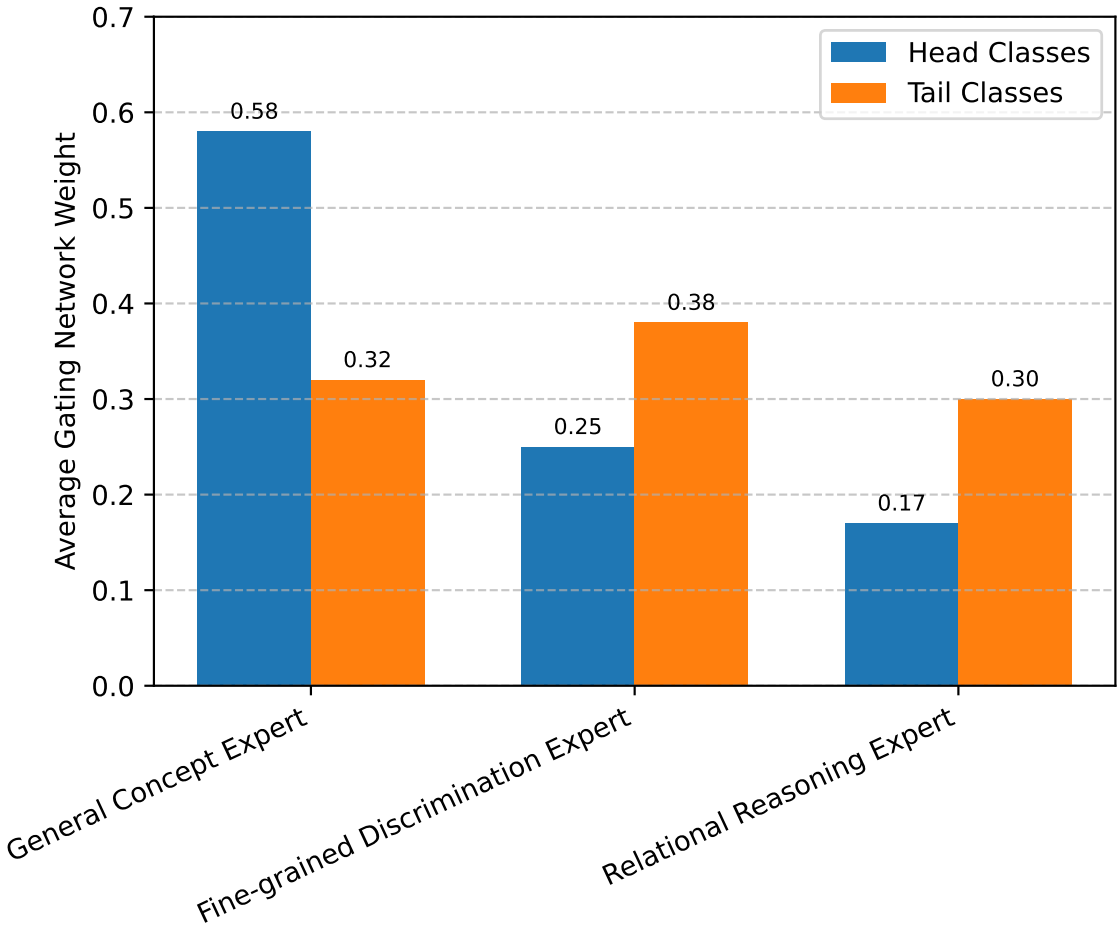

Figure 4. Average Gating Network Weights for Head and Tail Classes (Normalized, Sum to 1).

Furthermore, to quantify the performance gain attributable to this adaptive routing, we evaluate CARL-VU's performance on specific types of tail concepts that align with the strengths of E_F and E_R. For instance, for tail concepts requiring fine-grained distinctions (e.g., distinguishing specific dog breeds or subtle object attributes), and for those necessitating relational understanding (e.g., "person riding unicycle" vs. "person holding unicycle"), CARL-VU shows superior performance. This is summarized in Table 3, where we break down the VQAv2-LT accuracy for tail concepts based on their intrinsic complexity. The results demonstrate that the dynamic engagement of specialized experts by SER leads to tangible improvements in handling the diverse challenges presented by long-tailed data.

Table 3. VQAv2-LT Accuracy (%) for Tail Concepts Categorized by Semantic Complexity.

Method	Fine-grained Tail Concepts	Relational Tail Concepts
UniVLM (Base)	48.1	45.9
VL-Adapter	49.5	47.2
CARL-VU (Ours)	**53.8**	**52.1**

4.6. Impact of Contrastive Distillation (CD)

The Contrastive Distillation (CD) strategy is designed to explicitly enhance the discriminative power and compactness of feature representations for rare (tail) concepts. We analyze its effectiveness by examining performance variations with and without CD, and by tuning its key hyperparameter, the temperature τ.

Table 4 illustrates the impact of CD and different τ values on the performance of tail classes across VQA, captioning, and retrieval tasks. When CD is entirely removed (CARL-VU w/o CD), there is a

 doi:10.20944/preprints202511.1964.v1

notable drop in performance, particularly for tail-specific metrics. This confirms that explicitly refining the feature space for rare concepts is crucial. Furthermore, the choice of the temperature parameter τ plays a significant role. A value of $\tau = 0.07$ (as used in our final model) appears to strike an optimal balance, leading to the best performance. Lower τ values can make the contrastive loss too strict, potentially pushing apart genuinely similar samples, while higher τ values might make it too soft, failing to enforce sufficient discriminability. The improvement observed with CD, especially at an optimal τ, underscores its ability to create more robust and separable representations for tail classes.

Table 4. Effect of Contrastive Distillation and Temperature Parameter (τ) on Tail Class Performance.

CD Configuration	VQAv2-LT Tail	COCO-LT Tail	Flickr30K-LT Tail
CARL-VU w/o CD	49.3	75.8	33.5
CARL-VU w/ CD ($\tau = 0.05$)	50.7	77.1	34.8
CARL-VU w/ CD ($\tau = 0.07$)	**51.8**	**78.5**	**35.9**
CARL-VU w/ CD ($\tau = 0.10$)	50.9	77.4	35.0

To further quantify the effect of CD on the feature space, we measure the average intra-class similarity and inter-class dissimilarity for tail classes. Table 5 shows that with CD, tail classes exhibit higher average intra-class similarity (meaning samples of the same tail class are closer in the embedding space) and lower average inter-class similarity (meaning samples from different tail classes are further apart). This indicates that CD effectively distills more compact and separable feature clusters for rare concepts, directly addressing the feature collapse problem often seen in long-tailed learning.

Table 5. Average Feature Space Metrics for Tail Classes (Cosine Similarity).

CD Configuration	Avg. Intra-Class Similarity	Avg. Inter-Class Similarity
CARL-VU w/o CD	0.72	0.35
CARL-VU w/ CD ($\tau = 0.07$)	**0.81**	**0.28**

4.7. Effectiveness of Adaptive Feature Augmentation (AFA)

Adaptive Feature Augmentation (AFA) is crucial for mitigating data scarcity in tail classes by generating semantically consistent variations in the feature space. We analyze the individual and combined contributions of its core strategies: feature mixing and structured noise injection.

Table 6 presents the performance of CARL-VU when different AFA strategies are applied. The baseline "CARL-VU w/o AFA" shows a noticeable performance drop compared to the full model, highlighting the overall importance of AFA. When only Feature Mixing is employed, there is a substantial improvement, demonstrating its effectiveness in expanding the sample space and creating intermediate representations. Similarly, Structured Noise Injection alone also contributes positively, improving robustness by mimicking natural variations. The full AFA strategy, combining both Feature Mixing and Structured Noise Injection (along with linguistic augmentations), yields the best performance across all tail-specific metrics. This indicates a synergistic effect where the combination of diverse feature generation techniques provides the most comprehensive enrichment for scarce tail-class data.

 doi:10.20944/preprints202511.1964.v1

Table 6. Impact of Different Adaptive Feature Augmentation Strategies on Tail Class Performance.

AFA Strategy	VQAv2-LT Tail	COCO-LT Tail	Flickr30K-LT Tail
CARL-VU w/o AFA	48.7	74.9	33.1
CARL-VU w/ Feature Mixing only	50.1	76.5	34.2
CARL-VU w/ Structured Noise only	49.8	76.1	34.0
CARL-VU w/ Full AFA	**51.8**	**78.5**	**35.9**

The benefits of AFA are particularly pronounced for tail classes, as shown by their improved performance. By generating diverse, high-quality synthetic features, AFA effectively reduces the overfitting risk and enhances the model's generalization capabilities to unseen variations of rare concepts, directly addressing the core challenge of data scarcity.

4.8. Detailed Performance Analysis by Class Frequency

To provide a more granular understanding of CARL-VU's performance across different frequency strata, we present a detailed breakdown of results for head, medium, and tail classes. This analysis explicitly demonstrates CARL-VU's ability to boost performance on rare classes without compromising accuracy on common categories.

Table 7 compares CARL-VU with baseline models across head, medium, and tail class distributions for VQAv2-LT and COCO-LT. For VQAv2-LT, we categorize questions based on the frequency of their target answer concepts. For COCO-LT, concepts in captions are similarly categorized.

Table 7. Performance Breakdown by Class Frequency on VQAv2-LT (Acc. %) and COCO-LT (CIDEr).

Method	VQAv2-LT Accuracy (%)			COCO-LT CIDEr		
	Head	Medium	Tail	Head	Medium	Tail
UniVLM (Base)	62.1	56.8	47.9	85.2	79.5	72.1
VL-Adapter	62.5	57.3	48.5	84.8	79.1	72.5
CARL-VU (Ours)	**63.0**	**59.1**	**51.8**	**85.8**	**82.2**	**78.5**

As evidenced in Table 7, CARL-VU consistently outperforms baselines across all frequency bins, but its most significant gains are observed in the medium and tail categories. For VQAv2-LT, CARL-VU improves tail class accuracy by **+3.9%** over UniVLM (from 47.9% to 51.8%) and by **+3.3%** over VL-Adapter. Similarly, for COCO-LT, the improvement in tail CIDEr is substantial, showing a **+6.4** increase over UniVLM (from 72.1 to 78.5) and **+6.0** over VL-Adapter. Critically, CARL-VU also maintains competitive or slightly improved performance on head classes, demonstrating that its specialized mechanisms for long-tail learning do not negatively impact the understanding of common concepts. This balanced performance across the entire frequency spectrum highlights CARL-VU's robustness and its ability to effectively address the long-tail problem without sacrificing generalizability."

5. Conclusion

In this paper, we introduced Cross-Modal Adaptive Reasoning for Long-Tailed Visual-Linguistic Understanding (CARL-VU), a novel framework designed to tackle the pervasive challenge of long-tailed data distributions in Vision-Language Models (VLMs). CARL-VU addresses deep-rooted issues like feature bias and knowledge transfer through three core innovations within a robust Transformer-based Encoder-Decoder architecture: a semantic-driven expert routing (SER) mechanism for adaptive

 doi:10.20944/preprints202511.1964.v1

knowledge allocation, a contrastive distillation (CD) strategy to refine feature spaces for rare concepts, and adaptive feature augmentation (AFA) to enhance training diversity for tail categories. Our extensive experimental evaluation across VQAv2-LT, COCO-LT, and Flickr30K-LT unequivocally demonstrated CARL-VU's superior performance, consistently outperforming state-of-the-art baselines with significant gains, particularly in medium and tail categories, while critically maintaining or slightly enhancing head class performance. Comprehensive ablation studies confirmed the individual and synergistic contributions of SER, CD, and AFA, with SER showing the most significant impact, underscoring the importance of adaptive expert activation. Overall, CARL-VU represents a significant advancement towards building more robust and generalizable VLMs capable of effectively operating in real-world, long-tailed scenarios, laying a strong foundation for future research in complex reasoning and multimodal learning.

References

1. Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.
2. Zhitao Wang, Yirong Xiong, Roberto Horowitz, Yanke Wang, and Yuxing Han. Hybrid perception and equivariant diffusion for robust multi-node rebar tying. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 3164–3171. IEEE, 2025.
3. Zhihao Lin, Jianglin Lan, Christos Anagnostopoulos, Zhen Tian, and David Flynn. Safety-critical multi-agent mcts for mixed traffic coordination at unsignalized intersections. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2025.
4. Zhihao Lin, Jianglin Lan, Christos Anagnostopoulos, Zhen Tian, and David Flynn. Multi-agent monte carlo tree search for safe decision making at unsignalized intersections. 2025.
5. Zhitao Wang, Jiangtao Wen, and Yuxing Han. Ep-sam: An edge-detection prompt sam based efficient framework for ultra-low light video segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
6. Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738. Association for Computational Linguistics, 2022.
7. Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695. Association for Computational Linguistics, 2021.
8. Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.
9. Luqing Ren et al. Boosting algorithm optimization technology for ensemble learning in small sample fraud detection. *Academic Journal of Engineering and Technology Science*, 8(4):53–60, 2025.
10. Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886*, 2025.
11. Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548. Association for Computational Linguistics, 2022.
12. Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967. Association for Computational Linguistics, 2022.
13. Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100. Association for Computational Linguistics, 2022.
14. James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313. Association for Computational Linguistics, 2022.

 doi:10.20944/preprints202511.1964.v1

15. Zesheng Shi and Yucheng Zhou. Topic-selective graph network for topic-focused summarization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 247–259. Springer, 2023.
16. Yan Ling, Jianfei Yu, and Rui Xia. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159. Association for Computational Linguistics, 2022.
17. Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588. Association for Computational Linguistics, 2021.
18. Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775. Association for Computational Linguistics, 2022.
19. Feng Wang, Zesheng Shi, Bo Wang, Nan Wang, and Han Xiao. Readerlm-v2: Small language model for html to markdown and json. *arXiv preprint arXiv:2503.01151*, 2025.
20. Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. Less is more: Pretrain a strong Siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791. Association for Computational Linguistics, 2021.
21. Zesheng Shi, Yucheng Zhou, Jing Li, Yuxin Jin, Yu Li, Daojing He, Fangming Liu, Saleh Alharbi, Jun Yu, and Min Zhang. Safety alignment via constrained knowledge unlearning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25515–25529, 2025.
22. Luqing Ren. Reinforcement learning for prioritizing anti-money laundering case reviews based on dynamic risk assessment. *Journal of Economic Theory and Business Management*, 2(5):1–6, 2025.
23. Wang Jingzhi and Xuehao Cui. The impact of blood and urine biomarkers on age-related macular degeneration: insights from mendelian randomization and cross-sectional study from nhanes. *Biological Procedures Online*, 26(1):19, 2024.
24. Xuehao Cui, Dejia Wen, Jishan Xiao, and Xiaorong Li. The causal relationship and association between biomarkers, dietary intake, and diabetic retinopathy: insights from mendelian randomization and cross-sectional study. *Diabetes & Metabolism Journal*, 2025.
25. Zheng-Wei Liu, Jie Peng, Chun-Li Chen, Xue-Hao Cui, and Pei-Quan Zhao. Analysis of the etiologies, treatments and prognoses in children and adolescent vitreous hemorrhage. *International Journal of Ophthalmology*, 14(2):299, 2021.
26. ZQ Zhu, Peng Wang, NMA Freire, Ziad Azar, and Ximeng Wu. A novel rotor position-offset injection-based online parameter estimation of sensorless controlled surface-mounted pmsms. *IEEE Transactions on Energy Conversion*, 39(3):1930–1946, 2024.
27. Peng Wang, ZQ Zhu, and Dawei Liang. Improved position-offset based online parameter estimation of pmsms under constant and variable speed operations. *IEEE Transactions on Energy Conversion*, 39(2):1325–1340, 2024.
28. Xikun Wu, Mingyao Lin, Peng Wang, Lun Jia, and Xinghe Fu. Off-line stator resistance identification for pmsm with pulse signal injection avoiding the dead-time effect. In *2019 22nd International Conference on Electrical Machines and Systems (ICEMS)*, pages 1–5. IEEE, 2019.
29. Zhitao Wang, Weinuo Jiang, Wenkai Wu, and Shihong Wang. Reconstruction of complex network from time series data based on graph attention network and gumbel softmax. *International Journal of Modern Physics C*, 34(05):2350057, 2023.
30. Luqing Ren et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science*, 8(8):8–14, 2025.
31. Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161. Association for Computational Linguistics, 2021.
32. Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual*

 doi:10.20944/preprints202511.1964.v1

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2251–2262. Association for Computational Linguistics, 2022.

33. Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650. Association for Computational Linguistics, 2022.
34. Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630. Association for Computational Linguistics, 2021.
35. Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430. Association for Computational Linguistics, 2021.
36. Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540. Association for Computational Linguistics, 2021.
37. Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706. Association for Computational Linguistics, 2021.