LG-Cog: Language-Grounded Representation Learning with Cross-Modal Cognitive Priors for Image-Text Retrieval

1st Xiang Liu
School of Computer Science
Shenyang Aerospace University
Shenyang, China
liuxiang@stu.sau.edu.cn

2nd Yu Bai*
School of Computer Science
Shenyang Aerospace University
Shenyang, China
baiyu@sau.edu.cn

3rd Peng Lian
Shenyang Northern Software College
of Information Technology
Shenyang, China
445107168@qq.com

4th Haifeng Chi
School of Computer Science
Shenyang Aerospace University
Shenyang, China
chihaifeng@stu.sau.edu.cn

5th Haoxiang Li
School of Computer Science
Shenyang Aerospace University
Shenyang, China
lihaoxiang_sau@qq.com

***Abstract*—Contrastive learning has become a standard approach for image-text retrieval, aligning vision and language within a shared semantic space. However, in real-world scenarios, the assumed semantic consistency of positive pairs is frequently violated by noisy or partially matched data. To address this issue, we propose Language-Grounded Representation Learning with Cross-Modal Cognitive Priors (LG-Cog). LG-Cog leverages a frozen multimodal large language model with carefully designed prompts to generate fine-grained and semantically rich textual descriptions for each image, which serve as cross-modal cognitive priors complementary to the original paired texts. The joint alignment of image representations with the representations of both the original and generated texts drives visual representations toward a shared language semantic space, establishing a language-grounded representation learning paradigm. By integrating language-grounded representation learning with cross-modal cognitive priors, LG-Cog effectively mitigates semantic bias. Extensive experiments demonstrate the superiority of LG-Cog over state-of-the-art methods, achieving 3.7%–4.0% RSUM improvements on Flickr30K and MS-COCO benchmarks.**

***Index Terms*—image-text retrieval, language-grounded representation learning, cross-modal cognitive priors, multimodal large language model**

I. Introduction

With the rapid surge of multimodal data, understanding vision–language relationships has become crucial. Image–text retrieval is a pivotal task in this context, aiming to retrieve relevant images given text queries and relevant texts given image queries. Despite significant progress in image-text retrieval, mitigating the cross-modal semantic bias remains a major challenge for achieving accurate alignment.

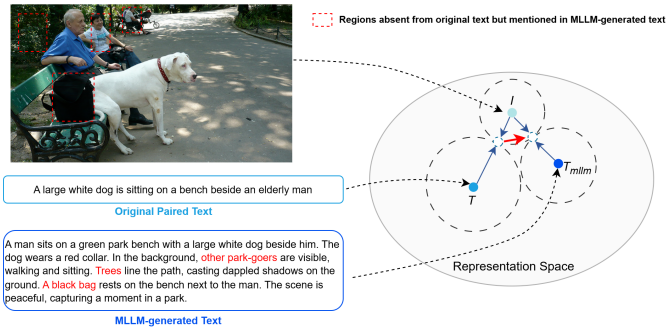

Fig. 1. Illustration of our motivation. Existing methods directly align image-text pairs but often suffer from semantic bias. To address this problem, we generate an additional semantically rich text for each image using a multimodal large language model (MLLM), which serves as a cross-modal cognitive prior complementary to the original paired text. The image representation is jointly aligned with the representations of both the original paired text and the generated text, driving it toward a shared language semantic space and progressively refining cross-modal alignment.

They optimize image and text embeddings using objectives such as hubness-aware loss [1] to pull semantically related pairs together and push mismatched pairs apart. This process can be seen as contrastive learning [2], with each image paired with its corresponding text as a positive pair, and non-matching texts treated as negatives. Contrastive learning assumes strict semantic consistency between positive pairs, but real-world data often violate this assumption. As illustrated in Fig. 1, even matched pairs can exhibit semantic bias—for instance, a text describing a large white dog beside an elderly man may omit visual elements like a *black bag* or *trees* present in the image. Ideally, the image representation I and the text representation T should be aligned to a shared semantic representation, but inherent discrepancies prevent this perfect alignment.

* Corresponding author.

This work was supported by the Fundamental Research Funds for the Universities of Liaoning Province(LJ232410143032).

To overcome this problem, we propose Language-Grounded Representation Learning with Cross-Modal Cognitive Priors (LG-Cog), a novel method for image-text retrieval. The core idea of LG-Cog is to encourage the image representation I to be jointly aligned with both the original text representation T and an additional semantically rich text representation T_{mllm} generated from the image, driving visual representations toward a shared language semantic space. Specifically, LG-Cog leverages a frozen multimodal large language model (MLLM) with carefully designed prompts to produce fine-grained and semantically rich textual descriptions conditioned on images, which serve as cross-modal cognitive priors complementary to the original paired texts. These prompts are designed to guide the MLLM toward producing objective and visually grounded content while avoiding subjective biases inherent to large language models. The joint alignment of image representations with the representations of both the original texts and the generated texts constitutes a language-grounded representation learning paradigm. By integrating language-grounded representation learning with cross-modal cognitive priors, LG-Cog effectively mitigates semantic bias.

Our main contributions can be summarized as follows:

- We propose LG-Cog, a novel image-text retrieval method that performs language-grounded representation learning by jointly aligning image representations with the representations of both the original paired texts and semantically enriched texts, driving visual representations toward a shared language semantic space.
- We systematically design and evaluate prompts for MLLM-based text generation, guiding the frozen multimodal large language model to produce objective, fine-grained, and semantically enriched descriptions that serve as cross-modal cognitive priors, thereby complementing the original paired texts.
- Extensive experiments on Flickr30K and MS-COCO show that LG-Cog outperforms state-of-the-art approaches by 3.7%–4.0% in terms of RSUM.

Our work is reproducible, and the code will be released soon.

II. Related Work

A. Image-Text Retrieval

Image-text retrieval methods can be primarily categorized into local-level matching and global-level matching.

Local-level matching methods perform detailed cross-modal interactions and align semantics at the level of local fragments, after which they compute an aggregate similarity score. Diao et al. [3] construct a method that enhances image-text matching by reasoning over similarity graphs and filtering out irrelevant alignments. Subsequently, Zhang et al. [4] design a negative-aware attention framework that leverages both matched and mismatched fragments to improve image-text similarity estimation, and Pan et al. [5] present a hard aligning network that focuses on the most relevant region-word pairs to improve accuracy and efficiency in image-text matching.

Global-level matching methods aim to learn holistic feature representations for images and texts, subsequently evaluating their similarity based on these global embeddings. Wang et al. [6] incorporate external commonsense knowledge into image-text matching by learning consensus-aware visual-semantic embeddings. Subsequently, Chen et al. [7] design a generalized pooling operator that automatically learns the optimal pooling strategy for visual-semantic embedding without manual tuning. Recently, Fu et al. [8] propose a hierarchical framework that models both fragment- and instance-level relationships to enhance cross-modal embeddings for image-text matching. Zhang et al. [9] devise a unified semantic enhancement framework with momentum contrast to improve the accuracy and efficiency of image-text retrieval.

B. Multimodal Large Language Models

Large Language Models (LLMs) have shown strong capabilities in language understanding, generation, and reasoning through large-scale pretraining on textual data. Representative models include T5 [10], which formulates diverse Natural Language Processing (NLP) tasks within a unified text-to-text framework, GPT-3 [11], which demonstrates powerful open-ended text generation via large-scale autoregressive modeling, and LLaMA [12], which emphasizes efficient and scalable language modeling with competitive performance. These models provide strong semantic abstraction and reasoning abilities, forming the foundation for subsequent extensions toward multimodal understanding.

Building upon LLMs, Multimodal Large Language Models (MLLMs) integrate visual encoders with language models to enable joint vision–language perception and reasoning. Early representative works include LLaVA [13], which aligns visual features with LLMs through visual instruction tuning to support multimodal dialogue. InternVL [14] further explores scalable multimodal pretraining with diverse vision–language data, while recent proprietary models such as GPT-4 [15] and Gemini [16] demonstrate strong general-purpose multimodal reasoning and perception capabilities across a wide range of tasks. Among open-source MLLMs, Qwen2.5-VL [17] exhibits strong fine-grained visual understanding and detailed visual-text generation abilities, making it particularly effective for producing semantically rich image descriptions. Overall, MLLMs achieve strong performance on multimodal understanding and generation tasks, including image captioning, visual question answering, and multimodal reasoning.

III. Method

As shown in Fig. 2, LG-Cog comprises a text generation module using a multimodal large language model, image encoders, text encoders, and dynamic queues. During training, the multimodal large language model with frozen parameters is leveraged to generate detailed textual descriptions for images. Both query and key encoders are utilized to generate embeddings for their respective modalities, with the key encoders' outputs being stored in their corresponding dynamic queues. During testing, only the Query Image Encoder and the Query

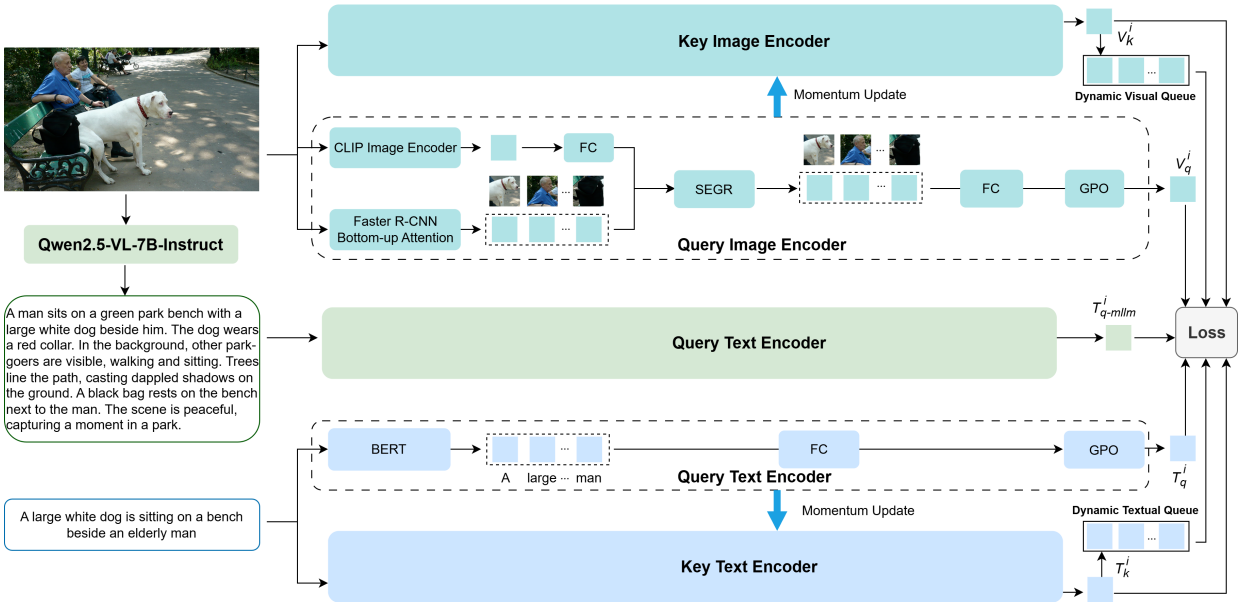

Fig. 2. The architecture of the proposed LG-Cog.

Text Encoder are employed to generate embeddings for image-text similarity computation; the multimodal large language model, key encoders, and dynamic queues are not used.

A. MLLM-Based Text Generation

We leverage a frozen multimodal large language model, Qwen2.5-VL-7B-Instruct [17], to generate detailed image descriptions that enrich textual representations for cross-modal retrieval. The multimodal large language model is prompted using the following instruction: *"Describe this image in the style of COCO or Flickr30K captions. Use concise, objective language. Focus on visual elements such as people, actions, objects, and background. Avoid opinions or interpretations. Use third-person perspective."*

Note that, to reduce computational overhead, the MLLM-based text generation is performed offline in advance.

The text generation process can be formulated as:

$$\text{Text}_{\text{mllm}} = \text{Qwen2.5-VL}(\text{Prompt}, \text{Image}). \tag{1}$$

For example, given an image with the original caption "A large white dog is sitting on a bench beside an elderly man," the generated text captures additional visual elements (e.g., "other park-goers", "trees", and "a black bag") that are absent from the original caption.

B. Image Encoders

To ensure feature consistency and support momentum updates, the Query and Key Image Encoders share the same architecture; thus, we describe the Query Image Encoder as a representative example.

Given an input image, we extract k=36 region features $V_r = \{v_i\}_{i=1}^{k}$, each with 2048 dimensions, using bottom-up attention [18]–[20] implemented with Faster R-CNN [21] and a 512-dimensional global feature v_c via CLIP (ViT-B/32) [22].

A Semantic Enhancement via Global Representation (SEGR) module enhances region-level features by modulating them with global semantic information. First, the global feature v_c is projected into the same feature space as the region features using a fully connected (FC) layer:

$$v_g = \text{FC}(v_c). \tag{2}$$

We compute attention weights between the global feature and each region feature as:

$$\lambda_i = \frac{\exp(v_g^T v_i)}{\sum_{j=1}^{k} \exp(v_g^T v_j)}. \tag{3}$$

The resulting attention weights are then used to modulate and enhance each region feature:

$$x_i = v_i + \lambda_i \cdot v_g. \tag{4}$$

$$V_{\text{r-enhance}} = \{x_i\}_{i=1}^{k}. \tag{5}$$

The enhanced region features are then projected and aggregated using the Generalized Pooling Operator (GPO) [7], a plug-and-play aggregation module that automatically consolidates features into holistic embeddings:

$$V = \text{GPO}(\text{FC}(V_{\text{r-enhance}})) \in \mathbb{R}^{d^J}, \tag{6}$$

where $d^J = 1024$ denotes the dimensionality of the shared image–text joint embedding space.

The outputs of the Query and Key Image Encoders for the i-th image are denoted as V_q^i and V_k^i, respectively.

C. Text Encoders

Similar to image encoders, the Query and Key Text Encoders share the same architecture; thus, we describe the Query Text Encoder as a representative example.

Given a text input of n tokens, we use BERT-base [23] to extract token-level features T_w, which are projected into the joint embedding space and aggregated via GPO [7] to obtain the final text embedding:

$$T = \text{GPO}(\text{FC}(T_w)) \in \mathbb{R}^{d^J}. \tag{7}$$

For the i-th original paired text, both Query and Key Text Encoders produce embeddings T_q^i and T_k^i. For the i-th MLLM-generated text, only the Query Text Encoder is used, producing $T_{q\text{-mllm}}^i$.

D. Cross-Modal Momentum Contrast

To stabilize contrastive learning, we adopt a momentum-based update strategy inspired by MoCo [24]. Specifically, the parameters of the query encoders are optimized via back-propagation, while those of the key encoders are updated using a momentum-based exponential moving average of the corresponding query encoder parameters:

$$\theta_{\text{key}}^V = m \cdot \theta_{\text{key}}^V + (1-m) \cdot \theta_{\text{query}}^V, \tag{8}$$

$$\theta_{\text{key}}^T = m \cdot \theta_{\text{key}}^T + (1-m) \cdot \theta_{\text{query}}^T, \tag{9}$$

where m denotes the momentum coefficient. Here, θ_{key}^V and θ_{key}^T denote the parameters of the Key Image Encoder and Key Text Encoder, respectively, while θ_{query}^V and θ_{query}^T denote the parameters of the Query Image Encoder and Query Text Encoder, respectively.

Dynamic visual and textual queues store embeddings from the key encoders, continuously enqueuing new embeddings and dequeuing the oldest, thus providing a large and diverse set of negative samples for contrastive learning.

E. Training Objectives

Our objective is to align matching image–text embeddings, push apart non-matching pairs, and align image embeddings with their corresponding MLLM-generated text embeddings.

We first adopt the hubness-aware loss (HAL) [1], which mitigates the hubness problem in high-dimensional embedding spaces by properly weighting positive and negative pairs, to align image-text representations:

$$\mathcal{L}_1 = \mathcal{L}_{\text{HAL}}(V_q, T_q) + \mathcal{L}_{\text{HAL}}(V_q, T_k) + \mathcal{L}_{\text{HAL}}(V_k, T_q), \tag{10}$$

where the first term uses positives and negatives from the current batch, while the latter two use query–key pairs as positives and obtain negatives from dynamic queues.

To align image embeddings with the MLLM-generated text embeddings, we extend the hubness-aware loss:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\log\left(1+\sum_{j\neq i} e^{\alpha(S_{ij}-\beta)}\right)}{\alpha} + \frac{\log\left(1+\sum_{j\neq i} e^{\alpha(S_{ji}-\beta)}\right)}{\alpha} - \log\left(1+\text{ReLU}(S_{ii})\right)\right), \tag{11}$$

where S_{ij} is the cosine similarity between the i-th image embedding V_q^i and the j-th MLLM-generated text embedding $T_{q\text{-mllm}}^j$, and S_{ii} denotes the similarity of the positive pair.

The overall loss integrates all objectives:

$$\mathcal{L} = \mathcal{L}_1 + \theta \cdot \mathcal{L}_2, \tag{12}$$

where θ weights the relative contribution of the alignment loss between images and MLLM-generated texts in the overall training objective.

IV. Experiments

A. Datasets and Implementation Details

We conduct experiments on two widely used benchmarks: Flickr30K [25] (31,000 images: 29,000 for training, 1,000 for validation, and 1,000 for testing) and MS-COCO [26] (123,287 images: 113,287 for training, 5,000 for validation, and 5,000 for testing), where each image is annotated with five captions. Following previous works [9], we evaluate the retrieval performance using R@K (K = 1, 5, 10), which measures the proportion of queries for which the correct item is ranked among the top K retrieved results. In addition, we report RSUM, defined as the sum of R@K values over the two retrieval directions, image-to-text and text-to-image, providing an aggregated measure of overall retrieval effectiveness.

All hyperparameters are determined based on validation performance. We set the temperature scale $\alpha = 90$, margin $\beta = 0.5$, and the momentum coefficient for the key encoders to $m = 0.999$. The loss weight θ is set to 0.7 for Flickr30K and 0.05 for MS-COCO. The dynamic queue size is 2048 for Flickr30K and 4096 for MS-COCO. We employ the AdamW optimizer [27] with weight decay 10^{-4}. For Flickr30K, we train the model for 20 epochs with an initial learning rate of 5×10^{-4}, which is decayed by a factor of 10 at epoch 10. For MS-COCO, the model is trained for 25 epochs with an initial learning rate of 4×10^{-4}, which is decayed by a factor of 10 at epoch 15. We conduct all experiments using PyTorch v1.11.0 on an NVIDIA A800 GPU.

B. Performance Comparison

We compare LG-Cog with state-of-the-art methods, including VSE∞ [7], VSRN++ [28], CHAN [5], CORA [29], USER [9], and CIMN [30]. LG-Cog is built upon the USER architecture [9].

We present comparative results on the Flickr30K and MS-COCO 1K test sets in Tables I and II. LG-Cog consistently achieves the best overall performance. On Flickr30K, LG-Cog outperforms USER, the strongest prior method, by 3.7% in RSUM and 1.6% for R@1 for image-to-text retrieval. On MS-COCO, LG-Cog surpasses CIMN, the previous best, by 4.0% in RSUM and 1.5% for R@1 for text-to-image retrieval.

C. Ablation Study

To verify the effectiveness of the proposed modules, we conduct ablation studies (Table III). Model I serves as conventional image-text alignment. Model II only introduces MLLM-based text generation, while Model III only adopts

TABLE I
COMPARISON RESULTS ON FLICKR30K TEST SET.

Method	Venue	Image-to-Text			Text-to-Image			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE∞ [7]	CVPR'21	81.7	95.4	97.6	61.4	85.9	91.5	513.5
VSRN++ [28]	TPAMI'22	79.2	94.6	97.5	60.6	85.6	91.4	508.9
CHAN [5]	CVPR'23	80.6	96.1	97.8	63.9	87.5	92.6	518.5
CORA [29]	CVPR'24	83.7	96.6	98.3	62.3	87.1	92.6	520.1
USER [9]	TIP'24	86.3	97.6	**99.4**	69.5	**91.0**	94.4	538.1
CIMN [30]	PR'25	84.4	96.1	98.2	64.8	88.1	92.8	524.4
LG-Cog (ours)	–	**87.9**	**98.1**	**99.4**	**70.5**	90.7	**95.2**	**541.8**

TABLE II
COMPARISON RESULTS ON MS-COCO 1K TEST SET.

Method	Venue	Image-to-Text			Text-to-Image			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE∞ [7]	CVPR'21	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++ [28]	TPAMI'22	77.9	96.0	98.5	64.1	91.0	96.1	523.6
CHAN [5]	CVPR'23	81.4	96.9	98.9	66.5	92.1	**96.7**	532.6
CORA [29]	CVPR'24	82.4	96.8	98.8	66.2	91.9	96.6	532.7
USER [9]	TIP'24	83.7	96.7	99.0	67.8	91.2	95.8	534.2
CIMN [30]	PR'25	82.2	97.0	99.0	67.4	**92.2**	96.6	534.4
LG-Cog (ours)	–	**85.1**	**97.3**	**99.1**	**68.9**	91.7	96.3	**538.4**

the cross-modal momentum contrast mechanism. Model IV combines both components. Compared to Model I, Model II improves overall retrieval performance, demonstrating that incorporating MLLM-generated semantically rich descriptions effectively enhances image-text alignment. Moreover, combining MLLM-based text generation with cross-modal momentum contrast (Model IV) consistently yields further gains over Model III, resulting in the best overall performance. These observations indicate that MLLM-based text generation is an effective strategy for improving image-text retrieval, and its combination with cross-modal momentum contrast produces complementary benefits.

TABLE III
ABLATION STUDY ON FLICKR30K.

Models	MLLM-Based Text Generation	Cross-Modal Momentum Contrast	Image-to-Text			Text-to-Image			RSUM
			R@1	R@5	R@10	R@1	R@5	R@10	
I			85.2	97.5	98.8	69.0	90.3	94.5	535.2
II	✓		84.9	97.9	99.2	69.4	90.9	94.4	536.7
III		✓	86.9	97.7	99.1	69.1	**91.2**	94.7	538.7
IV	✓	✓	**87.9**	**98.1**	**99.4**	**70.5**	90.7	**95.2**	**541.8**

TABLE IV
IMPACT OF DIFFERENT PROMPT DESIGNS ON FLICKR30K.

Prompts	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
Prompt I	86.9	97.7	99.3	69.5	**90.9**	94.9	539.2
Prompt II	87.3	**98.2**	**99.4**	69.4	90.7	94.7	539.7
Prompt III	**87.9**	98.1	**99.4**	**70.5**	90.7	**95.2**	**541.8**

We further analyze the impact of different prompt designs used for MLLM-based text generation, as summarized in Table IV.

Prompt I adopts a highly constrained captioning strategy by enforcing a single concise sentence: *"Describe the image with a single concise sentence in the style of COCO or Flickr30K image captions."* This design encourages the generated text to remain close to the distribution of the original dataset annotations.

Prompt II relaxes the sentence-number constraint while preserving the caption style: *"Describe this image in the style of COCO or Flickr30K captions."* By allowing multiple sentences, this prompt enables the model to include additional visual details beyond the original paired text.

Prompt III further refines Prompt II by explicitly emphasizing objective and observable visual content: *"Describe this image in the style of COCO or Flickr30K captions. Use concise, objective language. Focus on visual elements such as people, actions, objects, and background. Avoid opinions or interpretations. Use third-person perspective."* This design aims to suppress subjective bias and encourage fine-grained, visually grounded descriptions.

All three prompt designs improve image–text retrieval performance, with Prompt III yielding the largest gains. This demonstrates that guiding the MLLM to produce objective and fine-grained visual descriptions while minimizing its subjective bias is crucial.

D. Visualisation and Analysis

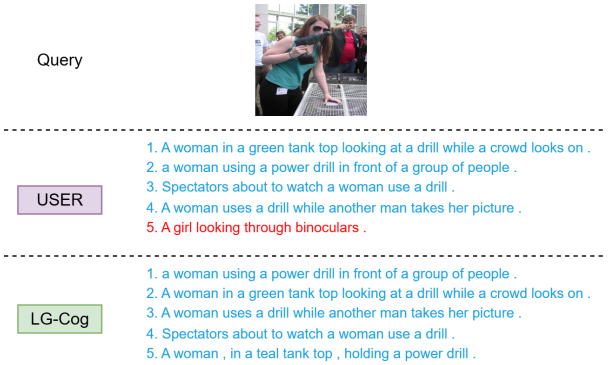

Fig. 3. Visualization of Image-to-Text retrieval results.

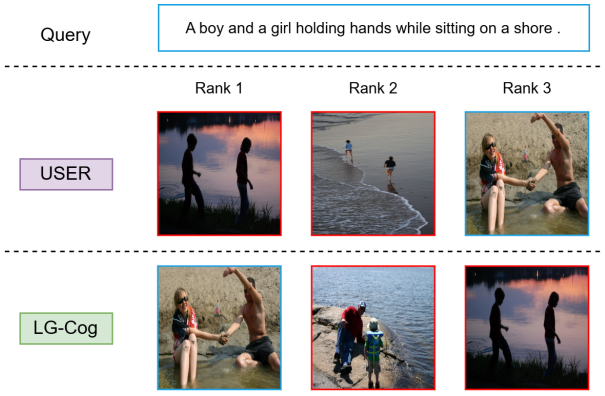

Fig. 4. Visualization of Text-to-Image retrieval results.

To validate LG-Cog's effectiveness, we compare its retrieval results with those of USER [9] in Fig. 3 and Fig. 4, showing top-5 image-to-text and top-3 text-to-image results (correct

results are highlighted in blue, incorrect results in red). USER often retrieves captions containing objects not present in the image (e.g., "binoculars") or returns top-1 images inconsistent with the text query (e.g., "standing" vs. "sitting"). In contrast, LG-Cog produces more accurate and semantically consistent image–text alignments across both tasks.

V. Conclusion

We propose LG-Cog, a novel image-text retrieval method designed to mitigate semantic bias arising from partially aligned image–text pairs in real-world data. LG-Cog employs a frozen multimodal large language model to generate semantically enriched texts for each image, which serve as cross-modal cognitive priors complementing the original texts. By jointly aligning image representations with both the original texts and the generated priors, LG-Cog drives visual representations toward a shared language semantic space, establishing a language-grounded representation learning paradigm that progressively refines cross-modal alignment. Extensive experiments on Flickr30K and MS-COCO demonstrate the superiority of LG-Cog for image-text retrieval.

References

[1] F. Liu, R. Ye, X. Wang, and S. Li, "Hal: Improved text-image matching by mitigating visual semantic hubs," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 563–11 571.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[3] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1218–1226.

[4] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 661–15 670.

[5] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 275–19 284.

[6] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 18–34.

[7] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 789–15 798.

[8] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, "Learning semantic relationship among instances for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 159–15 168.

[9] Y. Zhang, Z. Ji, D. Wang, Y. Pang, and X. Li, "User: Unified semantic enhancement with momentum contrast for image-text retrieval," *IEEE Transactions on Image Processing*, vol. 33, pp. 595–609, 2024.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[14] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 185–24 198.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[18] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[25] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[28] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Image-text embedding learning via visual and textual semantic reasoning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 641–656, 2022.

[29] K. Pham, C. Huynh, S.-N. Lim, and A. Shrivastava, "Composing object relations and attributes for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 354–14 363.

[30] X. Ke, B. Chen, X. Yang, Y. Cai, H. Liu, and W. Guo, "Cross-modal independent matching network for image-text retrieval," *Pattern Recognition*, vol. 159, p. 111096, 2025.