Benchmarking the Fine-Grained Discriminability in Image-Text Retrieval via Controlled Contrastive Differences

Zhen Wang[1,3], Xi Zhou[1,2,3,*], Yating Yang[1,2,3,*], Bo Ma[1,2,3], Lei Wang[1,2,3], Rui Dong[1,2,3], Azmat Anwar[1,2,3], Siru Miao[1,2,3],
[1]Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, China
{wang_zhen, zhouxi, yangyt, mabo, wanglei, dongrui, azmat}@ms.xjb.ac.cn

Abstract

Existing cross-modal image-text retrieval models often retrieve samples with inconsistent details. To evaluate fine-grained discriminability, we introduce MSCOCO-CCD and Flickr30k-CCD, with three key features: (1) a two-level image content taxonomy for contrastive sample generation and fine-grained evaluation; (2) annotation of numerous contrastive samples, where each sample differs from the anchor by a controlled contrastive difference (CCD), with the specific type of difference labeled; (3) a fine-grained contrastive discrimination metric to assess the ability to distinguish fine-grained nuances. Extensive experiments demonstrate that contrastive samples can significantly degrade retrieval performance. Furthermore, fine-grained evaluation reveals that current models still struggle to effectively produce discriminative representations on certain feature types, such as entity emotion and scene attribute. Our datasets and related codes will be publicly released [1].

1 Introduction

Cross-modal image-text retrieval (Wang et al., 2024b) aims to retrieve cross-modal target samples using text or image queries, which could be used in search engines and recommendation systems. With advancements in multimodal learning (Zeng et al., 2022; Li et al., 2023), particularly the multimodal large language models (MLLMs, Yin et al., 2023), retrieval models like VLM2VEC (Jiang et al., 2025) have demonstrated remarkable performance. However, as illustrated in Figure 1, they often retrieve samples with inconsistent details. Common benchmarks, such as MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) primarily assess coarse-grained retrieval performance. Using a retrieval model, Chen et al. (2023a) augments the candidate pool with overall similar samples, still unable to assess the discriminability across various types of fine-grained nuances.

Figure 1: Fine-grained retrieval across samples with controlled contrastive differences and our taxonomy.

To address this gap, we introduce MSCOCO-CCD and Flickr30k-CCD, designed to evaluate fine-grained retrieval and the discriminability on subtle differences. Our contributions are threefold: (1) A two-level image content taxonomy to guide contrastive sample generation and enable fine-grained evaluation by category, encompassing four primary dimensions: entity, scene, event/action, and style & presentation. (2) Using a Human-LLM collaborative annotation method, we annotate two large-scale datasets comprising contrastive samples, where each sample differs from the anchor by one controlled contrastive difference, with the specific type of difference labeled. (3) A fine-grained contrastive discrimination metric to evaluate the ability to capture subtle feature variations.

*Corresponding author
[1]https://github.com/miaomiao1215/CCD

Findings of the Association for Computational Linguistics: ACL 2026, pages 30491–30503

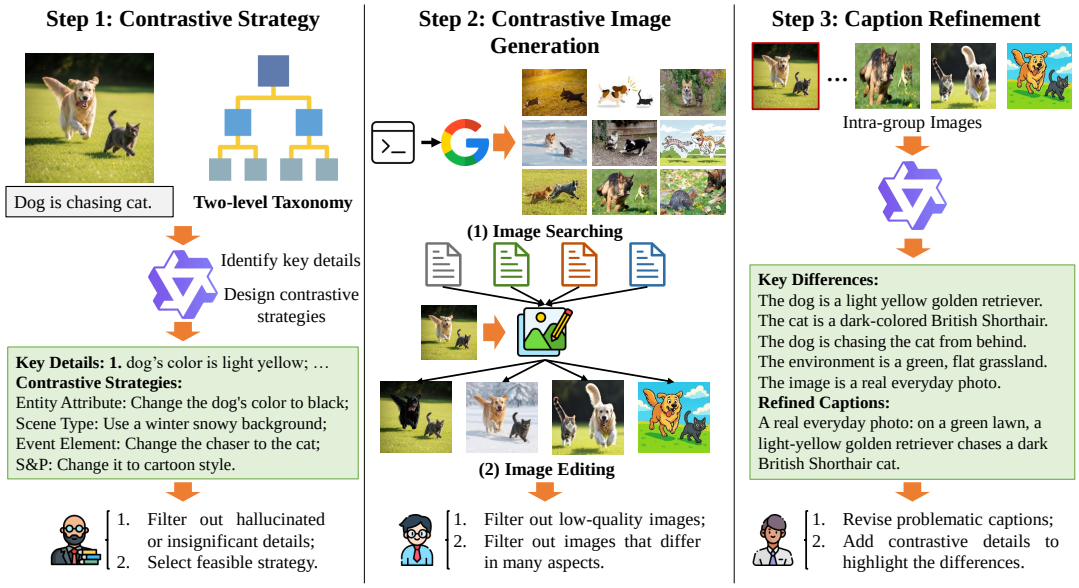

Figure 2: Illustration of the construction pipeline of our datasets.

Extensive experiments demonstrate the effectiveness of our datasets and that the introduction of contrastive samples leads to a notable decline in retrieval performance. Furthermore, fine-grained analysis reveals that existing models exhibit significant performance variations across feature types, showing a pronounced deficiency in more subtle features, such as entity emotion and scene attribute.

2 Method

Similar to existing benchmarks, CCDs encompass two sub-tasks: Text-to-Image (T2I) retrieval and Image-to-Text (I2T) retrieval. The dataset $D = \{d_i\}_{i=1}^{N}$ comprises N samples, each with one image I_i and five captions $T_i = \left\{t_i^j\right\}_{j=1}^{5}$. Notably, several challenging candidates with controlled contrastive differences in both modalities are incorporated. To facilitate a fine-grained evaluation, following a two-level taxonomy, each contrastive sample is annotated with a specific contrastive type.

2.1 Two-level Image Content Taxonomy

To establish a comprehensive evaluation framework, we develop a two-level taxonomy of image content, ensuring high coverage of common daily images. As illustrated in Figure 1 and Table 5, the taxonomy comprises four categories: Entity, Scene, Event/Action, and Style & Presentation. Each category is further subdivided into several secondary categories, such as entity type and emotion.

2.2 Dataset Creation

As illustrated in Figure 2, based on two widely-used benchmarks, Flickr30k and MSCOCO, we propose a Human-LLM collaborative annotation method to reduce annotation costs, which employs MLLMs to perform initial annotation, followed by rigorous human filtration and refinement.

Step 1: Contrastive Strategy. Given an image-caption pair, we input it into a powerful MLLM (Qwen3-VL-32B, Bai et al., 2025). Guided by our taxonomy and a few examples, the MLLM identifies key visual details and formulates several strategies for creating contrastive images. Each strategy includes a caption of the contrastive image, an image editing instruction, and a corresponding contrastive type. Subsequently, two human annotators verify the contrastive types, removing hallucinated details and infeasible strategies.

Step2: Contrastive Image Generation. We implement two image generation strategies: image search and image editing. Image search utilizes the captions of contrastive images as queries, retrieving relevant images via the Google Image Search API. A retrieval model is then employed to select the most similar images as potential candidates. The API's *as_rights* parameter is configured to ensure that only publicly available images are retrieved. Additionally, recent advances in image editing models enable the high-precision and controllable modifications. Therefore, image editing takes the anchor image and editing instructions as

Dim	MSCOCO		Flickr30k	
	Ori	Our	Ori	Our
Images	5000	19590	1000	3560
Avg.tokens	14.03	30.18	11.78	30.53
Avg.Sim_{image}	0.449	0.631	0.395	0.631
Avg.Sim_{text}	0.513	0.568	0.356	0.560

Table 1: Comparison of the original and our datasets.

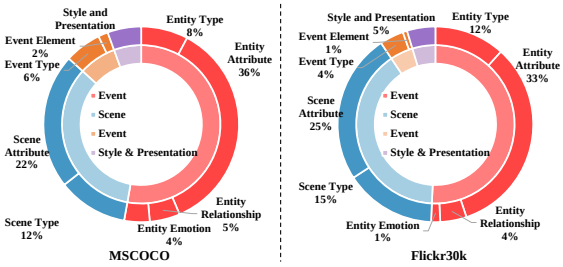

Figure 3: The distribution of the contrastive samples.

input to open-source Qwen-Image-Edit (Bai et al., 2025), yielding contrastive images that differ only in controlled details. Furthermore, as detailed in Appendix A.2 A.3, human annotators filter out low-quality images and those that deviate significantly from the anchor image.

Step3: Caption Refinement. The captions for each image must accurately reflect its visual content and highlight the differences from other images, incorporating at least one contrastive detail. As illustrated in Figure 1, a coarse-grained caption such as “A dog is chasing a cat” fails to distinguish subtle visual variations among similar candidates. These coarse-grained texts in existing benchmarks could lead to ambiguity problems where different non-target images all match the text, which is also proved in prior work "Rethinking Benchmarks for Cross-modal Image-text Retrieval". We prompt an MLLM with the same-group images to generate fine-grained captions adhering to these requirements. To ensure high-quality captions, annotators carefully review each image group and rectify any suboptimal captions.

Dataset Statistics. As shown in Table 1, our datasets are approximately four times larger than the original dataset. Additionally, the average token length of the captions is 2-3 times greater, providing more detailed descriptions. We further compute the mean cosine similarity between each sample and its top-5 noisy candidates. The results demonstrate that the top-5 candidates in our datasets exhibit higher similarity, thereby offering more challenging negative interference. As shown in Figure 3, *Entity* and *Scene* are the most prevalent categories, as most images contain these elements. In contrast, attributes such as "Emotion" are context-dependent and often absent in landscape or architectural photography. Despite the imbalance in category distribution, each category contains a sufficient number of samples to enable the evaluation of fine-grained discriminability. As demonstrated in Appendix 3, our datasets have been validated as high-quality, achieving an accuracy exceeding 95% in human evaluation.

2.3 Evaluation Metric

Following previous studies, we report Recall@1K ($K \in \{1, 5\}$) for all models, which measures the ratio of positive candidates being ranked in the top-K positions for all queries. Furthermore, we introduce a Fine-Grained Contrastive Discrimination Accuracy (FG-CDA) to evaluate the discriminability on subtle differences. Taking T2I as an example, given a pair of contrastive images I_i and I_j and a text query T_i^k (any caption of I_i and I_j), the testing model obtains their embeddings and computes their similarities. A retrieval is deemed successful only if the similarity with the target image $S(T_i^k, I_i)$ exceeds that of the contrastive image $S(T_i^k, I_j)$. The formula of FG-CDA is:

$$FG{-}CDA_{T2I} = \frac{\sum_{(i,j)} \sum_{k=1}^{5} I(S(T_i^k, I_i) > S(T_i^k, I_j)))}{5 \times Count((i,j))} \quad (1)$$

3 Experiments and Results

Models. To thoroughly assess the fine-grained retrieval performance of existing models, we conduct experiments on 11 CLIP-series models and 14 MLLM-based models. Specifically, CLIP-series models encompasses the original CLIP (Radford et al., 2021) and its variants: BLIP (Li et al., 2022), SigLIP (Zhai et al., 2023), AltCLIP (Chen et al., 2023b), and BGE (Zhou et al., 2025). For MLLM-based models, we include the VLM2Vec (Jiang et al., 2025), GME (Zhang et al., 2024), Ops-MM-embedding, UniME (Gu et al., 2025), BGE and RzenEmbed-v2 (Jian et al., 2025) families. More details are provided in Appendix B.2.

The Effectiveness of Our Benchmarks. We evaluate four models from different families across both T2I and I2T tasks, using the following dataset settings: (1) Ori: the original MSCOCO and Flickr30k (with coarse-grained captions and without con-

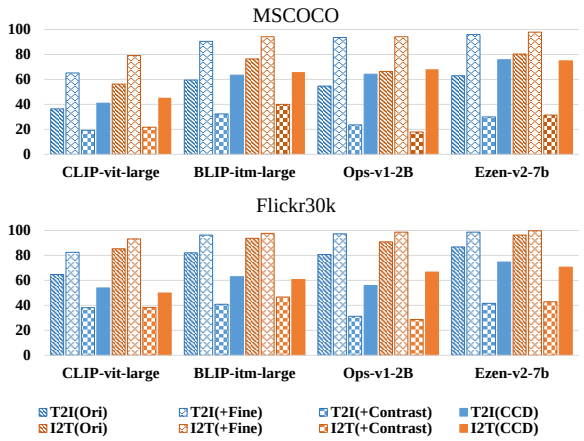

Figure 4: Comparison of R@1 between four different dataset settings on MSCOCO and Flickr30k.

trastive images); (2) + fine-grained captions: using our captions, but without contrastive images; (3) + contrastive images: adding contrastive images while retaining coarse-grained captions; (4) Our: MSCOCO-CCD and Flickr30k-CCD. As illustrated in Figure 4, when contrastive samples are added, a notable decline in R@1 is observed and the performance rankings across models are less consistent, with weaker CLIP-series models even surpassing SoTA MLLM-based models. Furthermore, applying fine-grained captions to the original datasets results in a significant improvement in R@1 for all models, particularly for the MLLM-based model, where R@1 approaches nearly 100%. In original datasets, coarse-grained captions often match multiple non-target images and underestimate retrieval performance. Therefore, to benchmark the actual retrieval preformance of retrieval models, caption refinement by incorporating at least one contrastive detail to reduce ambiguity in the text is necessary especially for the evaluation of fine-grained discriminability. Furthermore, when contrastive samples are added (from settings 1 to 3 and from 2 to 4), a notable decline in R@1 is observed, indicating a suboptimal ability to distinguish fine-grained differences. These results demonstrate that our CCDs are effective in evaluating fine-grained retrieval.

Overall Evaluation on CCDs. The experimental results for various models on MSCOCO-CCD are summarized in Table 2, where original samples serve as queries and both original and contrastive samples serve as candidates. Additional results on Flickr30k-CCD are provided in Table 4. We observe that all models exhibit relatively low

Model	Backbone	T2I		I2T	
		R1	R5	R1	R5
CLIP-series embedding models					
ALIGN-base	EfficientNet	46.7	83.6	50.3	83.0
CLIP	ViT-B(149M)	33.4	67.2	41.0	75.2
BGE-VL-base	ViT-B(149M)	56.2	88.8	40.4	69.9
BLIP-coco	ViT-B(149M)	65.8	94.7	64.0	92.5
BLIP-flickr30k	ViT-B(149M)	53.9	85.9	54.0	86.4
CLIP	ViT-L(307M)	41.1	75.6	45.4	79.9
BGE-VL-large	ViT-L(307M)	63.9	92.3	42.3	71.5
BLIP-coco	ViT-L(307M)	67.1	95.8	66.0	93.4
BLIP-flickr30k	ViT-L(307M)	59.7	**98.7**	61.9	96.6
SigLIP	SoViT(428M)	43.8	84.9	54.3	89.1
AltCLIP	ViT-H(632M)	43.5	81.2	47.6	80.0
MLLM-based embedding models					
GME	Qwen2-VL-2B	57.0	90.9	58.3	94.5
VLM2Vec	Qwen2-VL-2B	13.5	59.1	52.8	85.5
Ops-MM-E	Qwen2-VL-2B	64.2	96.2	68.2	93.8
UniME	Phi3.5-V-4.2B	64.7	93.9	64.9	91.6
VLM2Vec	LlaVA-1.6-7B	67.1	94.5	65.5	92.8
UniME	LlaVA-1.6-7B	69.6	94.4	68.1	94.4
UniME-OneV	LlaVA-1.6-7B	61.2	93.6	70.5	95.3
BGE-VL-S1	LlaVA-1.6-7B	62.6	94.1	65.4	90.9
BGE-VL-mmeb	LlaVA-1.6-7B	70.8	97.0	73.3	96.4
BGE-VL-S2	LlaVA-1.6-7B	73.3	97.1	69.8	94.7
VLM2Vec	Qwen2-VL-7B	31.4	82.9	63.1	91.3
GME	Qwen2-VL-7B	**77.7**	95.8	52.8	84.9
Ops-MM-E	Qwen2-VL-7B	70.5	97.2	71.7	95.0
Rzen-E	Qwen2-VL-7B	75.9	97.9	**75.0**	**96.9**

Table 2: Overall retrieval performance of various models on MSCOCO-CCD.

R@1 alongside higher R@5, especially the MLLM-based models, indicating that their retrieval performance is severely hindered by contrastive samples (averaging 2-3 for each anchor). Furthermore, MLLM-based models consistently outperform the CLIP series, with methods employing larger backbones demonstrating better performance. Notably, the Rzen-embed-v2-7b achieves the highest performance, likely due to its hardness-weighted mechanism for enhancing retrieval ability on challenging samples.

Evaluation on Fine-Grained Difference Discrimination. For each model, we compute the cosine similarities between embeddings of contrastive pairs from MSCOCO-CCD and Flickr30k-CCD and categorize the results using our taxonomy. As illustrated in Figure 5, certain categories, such as entity attribute and emotion, exhibit higher cosine similarities than scene types and S&P. Furthermore, the FG-CDA results for various models across different categories are depicted in Figure 6. Most models achieved high discrimination accuracies in scene types and S&P, even exceeding 95% accuracy in scene type. However, performance declines significantly on categories with more subtle fea-

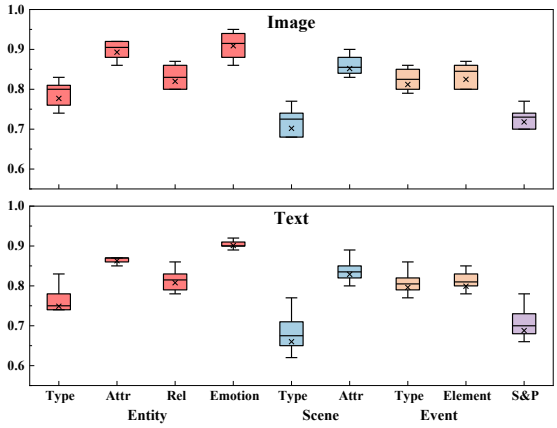

Figure 5: Comparison of cosine similarities between contrastive image pairs across 9 fine-grained categories.

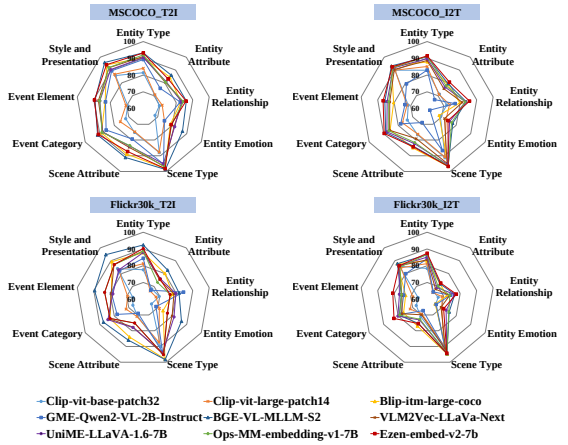

Figure 6: Comparison of FG-CDA across different fine-grained categories on nine LLMs.

tures, such as event attribute, entity emotion, and scene attribute. For example, I2T accuracy falls below 70% on entity emotion. This performance gap likely stems from the fact that emotions and attributes typically occupy smaller image regions. These findings shed light on future research: enhancing the feature capture and discriminability for more subtle aspects, such as attributes and emotions. Additionally, models with strong overall retrieval performance, such as Ezen-embed-v2-7b, also exhibit superior fine-grained discriminability.

4 Related Works

Benchmarks. MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) are two well-known image-text retrieval benchmarks, comprising large-scale image pool, each with five human-generated captions. Chen et al. (2023a) argue that the candidate pool and captions in MSCOCO an Flickr30k are coarse-grained, thereby introducing more overall similar images. Chen et al. (2023a) propose WIT, which is collected from Wikipedia and facilitates multilingual retrieval evaluation across 100+ languages. For the natural world domain, Vendrow et al. (2024) propose INQUIRE, which includes natural world images and expert-level retrieval queries. However, these datasets primarily assess overall retrieval performance and cannot systematically assess fine-grained retrieval and the ability to distinguish subtle differences.

Image-Text Retrieval Models Cross-modal image-text retrieval models primarily facilitate retrieval by leveraging unified representations of image and text features. CLIP (Radford et al., 2021) employs separate image and text encoders to extract features, followed by alignment through contrastive learning. CLIP-series models also include SigLIP (Zhai et al., 2023), BLIP (Li et al., 2022), AltCLIP (Chen et al., 2023b), etc. However, these models rely on late fusion of image and text features, thereby fail to capture deep relationships between them. In contrast, VLM2VEC (Jiang et al., 2025) introduces an MLLM-based framework that facilitates deep feature integration within a transformer architecture. Likewise, GME (Zhang et al., 2024), UniME (Gu et al., 2025) further improve it through multimodal data synthesis and hard negative sampling strategies.

5 Conclusion

We introduce MSCOCO-CCD and Flickr30k-CCD, two challenging fine-grained image-text retrieval benchmarks that incorporate extensive contrastive samples, each with one controlled contrastive difference from its anchor. Based on our taxonomy, these benchmarks evaluate the ability to discriminate subtle semantic nuances across diverse feature types. Experimental results demonstrate that even SoTA models exhibit significant performance degradation, showing there is room for developing fine-grained retrieval systems to capture subtle features and enhance fine-grained discriminability.

Acknowledgement

This research is sponsored by the Key Project of the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2025D01D45, 2023D01D17), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (2023B03024, 2024B03026), the Xin-

jiang Talent Development Fund(XJRC-2025-KJ-PY-KJLJ-048), the Major Scientific and Technological Projects of Xinjiang Uygur Autonomous Region(2025A03032), the Xinjiang Uygur Autonomous Region "Tianshan Talents" Innovation Team Project(No.2023TSYCTD0011), and the Youth Talents Support Project of Xinjiang Uyghur Autonomous Region (2023TSYCQNTJ0037).

Limitations

In this section, we discuss several limitations in our work. First, the construction of the dataset relies on high-performance image editing models and meticulous human annotation, which necessitates significant computational resources and human effort. Second, there is an absence of methods to enhance fine-grained discriminability. To address this, future work will explore strategies for automatically generating high-quality contrastive pairs and enhancing the capability to capture local features.

Ethics Statement

In this paper, we introduce two novel image-text retrieval datasets designed to assess fine-grained discriminability. First, image-text retrieval is an objective feature-matching task. Second, our CCDs are derived from two widely-used datasets, MSCOCO and Flickr30k, which contain verified everyday images. The augmented contrastive samples exhibit only fine-grained subtle differences from the anchor samples, ensuring no ethical violation. Note that we only used an AI assistant for writing polishing in the paper. AI was not involved in any other aspects of this research. In conclusion, our datasets avoid any ethical concerns.

License

To ensure copyright compliance and mitigate the risk of license infringement during dataset curation, we select images under licenses that permit open utilization and redistribution. The Google Search API is configured to filter for images either in the public domain or licensed under Creative Commons terms, with *rights* configured with "cc_publicdomain," "cc_attribute," "cc_sharealike," and "cc_noncommercial." Furthermore, we exclude images marked as "cc_nonderived". This method ensures that our data collection adheres to legal and ethical standards, respecting intellectual property rights while facilitating reproducible research.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Ibrahim M. Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Weijing Chen, Linli Yao, and Qin Jin. 2023a. Rethinking benchmarks for cross-modal image-text retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1241–1251. ACM.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023b. Altclip: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8666–8682. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the

modality barrier: Universal embedding learning with multimodal llms. *CoRR*, abs/2504.17432.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. Rzenembed: Towards comprehensive multimodal retrieval. *CoRR*, abs/2510.27350.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel J. Brostow, Kate E. Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. 2024. INQUIRE: A natural world text-to-image retrieval benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 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. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024b. Cross-modal retrieval: A systematic review of methods and future directions. *Proc. IEEE*, 112(11):1716–1754.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: improving universal multimodal retrieval by multimodal llms. *CoRR*, abs/2412.16855.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. 2025. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19076–19095. Association for Computational Linguistics.

Tester	T2I	I2T
A	96.5	97.8
B	95.8	97.5

Table 3: The human evaluation accuracies on two tasks.

A Dataset

A.1 Taxonomy

The definitions of the proposed two-level image content taxonomy are presented in Table 5 and are incorporated into the prompt template shown in Table 10 to support contrastive strategy generation.

A.2 Details in MLLM-based Annotation

As illustrated in Figure 2, the construction pipeline for our dataset consists of three distinct steps. In Step 1, we utilize Qwen-VL-32B-thinking-FP8 to extract key details from the provided images and implement several contrastive strategies. The associated prompt template is presented in Table 10, with a decoding temperature set to 0.0 to ensure consistency and reproducibility. To optimize computational efficiency, images are resized while maintaining their original aspect ratio, with pixel counts constrained between 256*784 and 840*504 pixels. In the step 2, the API's as_rights parameter of the Google Image Search API is configured to ensure that only publicly available images are retrieved. After the image search, the top five retrieved images corresponding to the query are embedded using Ops-MM-embedding-7B, with cosine similarity scores employed to identify the most similar images as potential candidates. Additionally, to keep a balance between computational cost and performance, the inference steps for Qwen-Image-Edit are set to 28. In Step 3, using the prompt template outlined in Table 10, Qwen-VL-32B-thinking-FP8 is employed to refine the captions of images with the decoding temperature of 0.0.

A.3 Human Annotator

Our annotation team consists of two graduate students engaged in multimodal image-text retrieval. They are trained and required to annotate the data according to the predefined guidelines in Section 2.2. Another annotator reviews their annotation results and eliminates their discrepancies in the annotation of contrastive category. Any discrepancies during the review process are discussed and resolved to ensure high-quality and reliable annotations.

Model	Backbone	T2I		I2T	
		R1	R5	R1	R5
CLIP-series embedding models					
CLIP	ViT-B(149M)	46.9	86.4	48.0	89.3
BGE-VL-base	ViT-B(149M)	66.0	97.7	45.5	81.6
BLIP-coco	ViT-B(149M)	58.1	94.4	58.7	93.6
BLIP-flickr30k	ViT-B(149M)	62.1	96.5	56.3	92.2
ALIGN-base	EfficientNet	53.9	95.0	54.8	92.2
CLIP	ViT-L(307M)	53.9	91.4	49.9	89.9
BGE-VL-large	ViT-L(307M)	61.7	95.8	49.8	84.4
BLIP-coco	ViT-L(307M)	63.0	98.3	60.7	94.2
BLIP-flickr30k	ViT-L(307M)	63.5	94.7	63.5	93.1
SigLIP	SoViT(428M)	49.2	94.4	57.1	94.7
AltCLIP	ViT-H(632M)	49.1	93.4	51.0	90.2
MLLM-based embedding models					
GME	Qwen2-VL-2B	41.9	93.9	58.3	94.5
VLM2Vec	Qwen2-VL-2B	11.2	76.1	51.6	92.7
Ops-MM-E	Qwen2-VL-2B	55.9	98.6	66.7	96.3
UniME	Phi3.5-V-4.2B	63.5	97.8	62.9	95.4
VLM2Vec	LlaVA-1.6-7B	66.4	98.3	60.6	96.1
UniME	LlaVA-1.6-7B	71.0	98.1	64.1	96.5
UniME-OneV	LlaVA-1.6-7B	59.4	98.5	67.9	97.0
BGE-VL-S1	LlaVA-1.6-7B	62.2	97.6	64.4	94.6
BGE-VL-mmeb	LlaVA-1.6-7B	69.0	99.2	67.7	**97.8**
BGE-VL-S2	LlaVA-1.6-7B	73.6	99.4	65.4	96.6
VLM2Vec	Qwen2-VL-7B	32.1	93.0	61.6	97.1
GME	Qwen2-VL-7B	65.8	97.7	55.1	95.3
Ops-MM-E	Qwen2-VL-7B	70.6	99.4	66.5	**97.7**
Rzen-E	Qwen2-VL-7B	**74.7**	**99.5**	**70.6**	97.1

Table 4: Overall retrieval performances of various retrieval models on Flickr30k-CCD.

A.4 The Quality of the Annotation

To further assess the quality of the datasets, we conducted a human evaluation involving two independent testers who were not involved in the data annotation. Specifically, we merged two datasets and randomly selected 400 samples, along with their top-5 most similar non-target samples (Ops-MM-embedding-v1-7B) to serve as noise candidates. In the I2T task, the testers were asked to select the target caption from six text candidates based on the provided image. For the T2I task, the testers were asked to select the target image from six image candidates based on the provided caption. As shown in Table 3, both testers achieved over 95% accuracy in both tasks, thereby validating the effectiveness of the dataset's annotations.

B Experiments

B.1 Implementation Details

Experiments are conducted on four NVIDIA Tesla A100 GPUs with 80GB of RAM each. To ensure a fair comparison, the models are assessed using their default configuration. Additionally, MLLM-based

Type	Definition
Entity Type	The specific fine-grained category of the entity, such as Donald Trump, Labrador, Husky, Boeing 747, etc.
Entity Attribute	The specific visual or intrinsic characteristics of an entity, encompassing aspects, such as color, pattern, material, shape, occupation, etc.
Entity Relation	Spatial, interactive, or relational dynamics among entities, such as on top of, riding, shaking hands, husband and wife, friends, etc.
Entity Emotion	Emotions of the entity, such as positive, negative, sad, and happy.
Scene Type	The fine-grained categorization of the scene, such as market, lakeside, seaside, church interior, etc.
Scene Attribute	Visual features or the state of the scene, such as crowded market, tranquil lakeside, clear blue sky, etc.
Event/ Action Type	Specific types of actions or events, such as chasing, dancing, and caressing.
Event/ Action Attribute	Participants and characters that constitute the action or event, such as subject and object.
Style & Presentation	Specific mode of visual representation and artistic expression, such as photograph, painting, illustration, comic, ink painting, impressionism, realism, etc.

Table 5: The definitions of our two-level image content taxonomy.

models, such as VLM2VEC, utilize the prompts specified in their papers. For example, the I2T and T2I tasks for VLM2VEC emplot the instructions *Find an image caption describing the given everyday image.* and *Find me an everyday image that matches the given caption.*, respectively.

B.2 Models in Experiments

In this study, we conduct a comprehensive evaluation of the fine-grained retrieval capabilities across a diverse set of models. These models leverage backbones with varying parameter sizes, including ViT (base, large, and huge, Dosovitskiy et al., 2021), SoViT (Alabdulmohsin et al., 2023), Qwen2-VL (2B and 7B, Wang et al., 2024a), Phi3.5-V-4.2B (Abdin et al., 2024), and LlaVA-1.6-7B (Liu et al., 2023). The specific models utilized in our experiments are outlined below:

- *CLIP*[2] employs a Vision Transformer (ViT) for encoding images and utilizes a Transformer model (Vaswani et al., 2017) for processing text. The alignment between the two modalities is then achieved through contrastive learning on extensive image-text datasets.
- *BLIP*[3] introduces a captioning model designed to generate synthetic captions for web data, as well as a filtering method to eliminate noisy captions.
- *ALIGN*[4] employ EfficientNet (Tan and Le, 2019) as its vision encoder and BERT (Devlin et al., 2019) as its text encoder, employing contrastive learning on a vast noisy dataset.
- *SigLIP*[5] replace softmax normalization with sigmoid loss in contrastive learning, enabling larger batch size and improving training efficiency.
- *AltCLIP*[6] is developed on Stable Diffusiosn and trained on parallel corpus, providing an enhanced bilingual CLIP model.
- *VLM2VEC*[7] introduces a unified MLLM-based framework that facilitates deep feature integration within a transformer architecture, which is suitable for a wide range of tasks.
- *BGE-VL*[8] introduce a novel data synthesis method, MegaPairs, which leads to significant improvements in performance.
- *GME*[9] develop a pipeline for synthesizing training data and creates a large-scale, fused-modal dataset to address the modality imbalance.
- *UniME*[10] develop a novel two-stage framework

[2] https://huggingface.co/openai/clip-vit-base-patch32, https://huggingface.co/openai/clip-vit-large-patch14

[3] https://huggingface.co/Salesforce/blip-itm-base-coco, https://huggingface.co/Salesforce/blip-itm-base-flickr, https://huggingface.co/Salesforce/blip-itm-large-coco, https://huggingface.co/Salesforce/blip-itm-large-flickr

[4] https://huggingface.co/kakaobrain/align-base

[5] https://huggingface.co/google/siglip-so400m-patch14-384

[6] https://huggingface.co/BAAI/AltCLIP

[7] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-2B, https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-7B, https://huggingface.co/TIGER-Lab/VLM2Vec-LLaVa-Next

[8] https://huggingface.co/BAAI/BGE-VL-base, https://huggingface.co/BAAI/BGE-VL-large, https://huggingface.co/BAAI/BGE-VL-MLLM-S1, https://huggingface.co/BAAI/BGE-VL-v1.5-mmeb, https://huggingface.co/BAAI/BGE-VL-MLLM-S2

[9] https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-7B-Instruct, https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-2B-Instruct

[10] https://huggingface.co/DeepGlint-AI/UniME-Phi3.5-V-4.2B, https://huggingface.co/DeepGlint-AI/UniME-LLaVA-1.6-7B, https://huggingface.co/DeepGlint-AI/UniME-LLaVA-OneVision-7B

that leverages MLLMs to learn discriminative representations, which performs textual discriminative knowledge distillation from a powerful LLM-based teacher and introduces a hard negative enhanced instruction tuning to further advance discriminative representation learning.

- *Ops-MM-embedding-v1*[11] is fine-tuned from Qwen2-VL, which encodes text, images, text-image pairs, visual documents, and videos into a unified embedding space for cross-modal retrieval.
- *RzenEmbed-v2-7B*[12] learn embeddings across a diverse set of modalities, including text, images, videos, and visual documents, and employ a novel two-stage training strategy to learn discriminative representations.

[11] https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B, https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-7B

[12] https://huggingface.co/qihoo360/RzenEmbed

Model	Backbone	Text to Image											S&P
		Entity					Scene			Event			
		Type	Attr	Rel	Emo	Avg	Type	Attr	Avg	Event	Event	Avg	
CLIP-series embedding models													
CLIP	ViT-B(149M)	18.6	29.4	30.2	31.9	27.1	12.0	24.8	20.0	27.9	30.7	28.3	14.3
BGE-VL-base	ViT-B(149M)	10.7	15.8	19.4	13.4	14.9	2.5	10.3	7.3	16.9	19.4	17.3	4.7
BLIP-coco	ViT-B(149M)	10.9	16.6	16.7	19.8	15.4	2.4	12.7	8.8	11.9	17.0	12.6	6.8
BLIP-flickr30k	ViT-B(149M)	14.9	21.3	25.0	23.6	20.2	5.8	19.1	14.0	20.6	25.3	21.3	10.3
ALIGN-base	EfficientNet+BERT	13.8	23.7	23.8	21.0	21.3	7.1	20.0	15.1	19.1	22.7	19.7	10.7
CLIP	ViT-L(307M)	16.0	29.2	28.5	25.4	26.0	12.3	25.3	20.3	24.2	29.5	25.0	13.3
BGE-VL-large	ViT-L(307M)	9.0	14.6	16.9	11.3	13.4	1.6	9.1	6.3	14.7	14.6	14.7	3.9
BLIP-coco	ViT-L(307M)	9.6	15.4	17.2	21.3	14.4	1.9	11.1	7.6	11.2	16.9	12.0	6.0
BLIP-flickr30k	ViT-L(307M)	10.6	16.3	18.3	23.7	15.4	2.2	12.1	8.4	13.7	18.4	14.4	6.8
SigLIP	SoViT(428M)	11.3	19.5	22.9	18.7	17.9	7.3	20.4	15.5	15.6	25.3	17.1	10.0
AltCLIP	ViT-H(632M)	13.9	25.4	24.2	21.5	22.5	8.2	21.6	16.5	20.5	25.0	21.2	9.7
MLLM-based embedding models													
GME	Qwen2-VL-2B	10.7	24.3	17.6	25.4	20.6	4.7	20.7	14.6	14.4	17.1	14.8	10.6
VLM2Vec	Qwen2-VL-2B	20.8	36.4	28.1	41.6	32.3	14.0	35.6	27.4	28.3	31.2	28.7	18.8
Ops-MM-E	Qwen2-VL-2B	7.8	19.5	16.7	19.7	16.6	3.3	16.7	11.7	11.3	15.7	12.0	8.0
UniME	Phi3.5-V-4.2B	9.9	19.0	15.8	14.4	16.5	2.7	14.0	9.7	10.6	14.1	11.1	6.8
VLM2Vec	LlaVA-1.6-7B	8.9	19.5	16.0	18.7	16.8	3.4	15.4	10.8	12.1	12.8	12.2	8.2
UniME	LlaVA-1.6-7B	9.6	18.9	16.8	18.3	16.6	4.9	16.1	11.9	11.9	13.7	12.2	9.6
UniME-OneVision	LlaVA-1.6-7B	8.9	19.8	16.5	21.1	17.1	3.0	14.0	9.8	13.4	14.3	13.5	6.7
BGE-VL-MLLM-S1	LlaVA-1.6-7B	9.4	16.2	20.3	15.7	15.0	2.5	12.0	8.4	14.0	18.5	14.7	6.0
BGE-VL-v1.5-mmeb	LlaVA-1.6-7B	7.5	14.4	14.5	14.9	12.9	1.7	10.1	6.9	9.6	12.2	10.0	5.2
BGE-VL-MLLM-S2	LlaVA-1.6-7B	6.7	14.0	13.9	13.0	12.3	1.3	8.9	6.1	8.5	10.8	8.9	4.1
VLM2Vec	Qwen2-VL-7B	15.4	32.4	23.5	33.0	27.8	6.9	25.1	18.3	20.9	21.6	21.0	11.6
GME	Qwen2-VL-7B	12.0	25.2	19.5	27.5	21.7	4.7	22.3	15.6	13.9	18.4	14.6	12.9
Ops-MM-E	Qwen2-VL-7B	8.0	19.4	15.7	20.5	16.5	2.9	16.5	11.3	11.6	13.6	11.9	8.1
Rzen-E	Qwen2-VL-7B	6.8	16.9	14.1	20.8	14.5	1.8	12.7	8.6	9.0	10.4	9.3	5.8

Table 6: The Fine-Grained Contrastive Discrimination Error (FG-CDE) results for various models on the Text-to-Image (T2I) task of MSCOCO-CCD. Note that the sum of FG-CDE and FG-CDA equals 100%.

Model	Backbone	Text to Image											S&P
		Entity					Scene			Event			
		Type	Attr	Rel	Emo	Avg	Type	Attr	Avg	Event	Event	Avg	
CLIP-series embedding models													
CLIP	ViT-B(149M)	17.1	28.7	28.1	30.9	26.0	8.8	23.4	17.8	26.3	28.5	26.7	10.1
BGE-VL-base	ViT-B(149M)	18.5	27.2	27.3	28.1	25.2	7.6	20.5	15.6	25.6	27.0	25.8	9.7
BLIP-coco	ViT-B(149M)	12.4	21.7	19.6	28.5	19.6	3.4	15.9	11.1	15.1	19.7	15.8	7.4
BLIP-flickr30k	ViT-B(149M)	13.9	23.6	23.7	30.4	21.6	5.3	19.4	14.1	19.3	23.7	20.0	9.5
ALIGN-base	EfficientNet+BERT	14.0	26.3	23.5	26.3	23.2	6.1	21.2	15.5	20.0	23.4	20.5	10.0
CLIP	ViT-L(307M)	14.9	29.0	26.4	27.0	25.5	9.3	23.7	18.2	22.9	27.8	23.7	10.1
BGE-VL-large	ViT-L(307M)	17.5	24.5	26.9	26.8	25.1	7.1	20.5	15.4	24.5	26.0	24.8	10.4
BLIP-coco	ViT-L(307M)	12.2	21.1	20.2	31.8	19.3	3.2	15.0	10.5	14.5	19.7	15.3	6.9
BLIP-flickr30k	ViT-L(307M)	11.9	20.9	21.1	32.0	19.2	3.2	15.3	10.7	16.0	21.1	16.7	7.3
SigLIP	SoViT(428M)	13.1	24.6	25.4	28.6	22.1	7.7	22.4	16.8	19.6	26.8	20.7	10.0
AltCLIP	ViT-H(632M)	13.6	27.5	23.5	27.1	23.9	8.0	22.1	16.7	20.3	25.3	21.0	8.6
MLLM-based embedding models													
GME	Qwen2-VL-2B	17.0	33.1	23.1	38.1	28.7	13.3	31.1	24.4	21.8	26.5	22.5	20.8
VLM2Vec	Qwen2-VL-2B	13.4	27.9	22.5	30.4	24.2	7.5	24.9	18.3	17.9	23.1	18.7	11.2
Ops-MM-E	Qwen2-VL-2B	10.2	24.3	18.5	26.7	20.6	5.4	20.9	15.0	16.1	18.5	16.4	9.1
UniME	Phi3.5-V-4.2B	10.8	23.3	17.6	25.1	20.0	4.0	18.2	12.9	12.6	15.5	13.1	7.9
VLM2Vec	LlaVA-1.6-7B	11.3	24.6	18.7	28.1	21.1	5.2	20.9	15.0	14.1	17.3	14.6	9.5
UniME	LlaVA-1.6-7B	10.2	23.3	16.7	24.5	19.7	4.4	17.7	12.7	12.2	15.7	12.7	8.4
UniME-OneVision	LlaVA-1.6-7B	8.4	21.1	14.7	21.4	17.6	3.3	16.1	11.3	10.3	14.2	10.9	6.7
BGE-VL-MLLM-S1	LlaVA-1.6-7B	13.3	23.9	22.2	27.3	21.4	3.7	15.9	11.3	18.2	22.1	18.8	7.7
BGE-VL-v1.5-mmeb	LlaVA-1.6-7B	8.6	19.8	15.6	22.3	16.9	3.0	15.4	10.7	11.1	14.5	11.6	7.3
BGE-VL-MLLM-S2	LlaVA-1.6-7B	9.1	20.6	16.4	22.4	17.7	3.4	16.0	11.3	11.4	14.3	11.9	6.8
VLM2Vec	Qwen2-VL-7B	11.0	25.5	18.1	26.9	21.6	4.9	20.2	14.4	13.7	17.9	14.3	8.6
GME	Qwen2-VL-7B	18.4	35.6	23.9	40.6	30.8	16.6	35.3	28.2	22.8	27.0	23.5	25.9
Ops-MM-E	Qwen2-VL-7B	8.8	22.3	16.9	23.2	18.8	3.4	18.4	12.7	13.5	16.7	14.0	8.2
Rzen-E	Qwen2-VL-7B	8.6	19.4	14.3	25.7	16.6	3.2	15.7	11.0	10.3	13.3	10.8	7.6

Table 7: The FG-CDE results for various models on the Image-to-Text (I2T) task of MSCOCO-CCD.

Model	Backbone	Text to Image											S&P
		Entity					Scene			Event			
		Type	Attr	Rel	Emo	Avg	Type	Attr	Avg	Event	Event	Avg	
CLIP-series embedding models													
CLIP	ViT-B(149M)	22.4	32.1	30.3	34.3	30.7	9.7	29.5	22.8	32.9	33.7	33.0	16.1
BGE-VL-base	ViT-B(149M)	13.7	20.8	23.1	17.7	19.8	1.9	15.1	10.6	24.3	21.3	23.7	7.3
BLIP-coco	ViT-B(149M)	15.5	21.9	23.8	25.2	21.4	2.8	18.6	13.2	19.8	20.0	19.8	12.0
BLIP-flickr30k	ViT-B(149M)	15.1	21.7	25.4	26.4	21.5	5.6	23.9	17.6	22.1	23.7	22.4	15.6
Align-base	EfficientNet+BERT	15.4	27.0	24.9	25.8	25.0	5.2	24.9	18.1	24.0	32.4	25.8	16.9
CLIP	ViT-L(307M)	18.7	33.6	32.3	28.9	30.9	12.6	29.0	23.4	28.2	30.5	28.7	17.3
BGE-VL-large	ViT-L(307M)	11.2	19.3	21.2	14.0	17.9	1.7	13.9	19.7	20.5	16.8	19.7	7.5
BLIP-coco	ViT-L(307M)	12.2	19.9	22.2	26.4	19.5	2.1	16.0	11.2	17.0	20.8	17.8	10.8
BLIP-flickr30k	ViT-L(307M)	11.9	19.0	23.6	27.3	19.1	2.4	16.9	12.0	18.3	22.1	19.1	13.1
SigLIP	SoViT(428M)	12.0	22.3	25.7	22.1	21.1	7.2	24.7	18.7	19.7	21.3	20.0	14.1
AltCLIP	ViT-H(632M)	14.7	30.1	27.9	25.1	27.2	8.2	25.3	19.5	25.7	27.6	26.1	14.4
MLLM-based embedding models													
GME	Qwen2-VL-2B	15.6	32.9	15.6	31.2	29.6	6.3	31.0	22.6	22.0	20.8	21.7	17.9
VLM2Vec	Qwen2-VL-2B	28.3	41.9	34.3	40.1	39.0	16.3	41.6	32.9	35.9	33.7	35.5	25.3
Ops-MM-E	Qwen2-VL-2B	10.2	26.9	23.2	24.0	23.9	4.8	24.5	17.8	18.4	19.5	18.6	13.7
UniME	Phi3.5-V-4.2B	13.6	24.9	22.1	18.4	22.5	4.1	22.4	16.2	19.0	19.0	19.0	12.0
VLM2Vec	LlaVA-1.6-7B	11.2	24.8	21.4	20.1	22.1	5.2	22.2	16.4	16.6	20.5	17.5	12.8
UniME	LlaVA-1.6-7B	12.1	24.1	20.3	18.8	21.5	6.0	22.0	16.6	15.6	21.6	16.9	17.2
UniME-OneVision	LlaVA-1.6-7B	10.9	24.7	20.7	22.5	22.1	4.5	20.5	15.0	17.5	20.3	18.1	10.2
BGE-VL-MLLM-S1	LlaVA-1.6-7B	12.7	20.5	27.0	17.6	19.7	2.3	16.3	11.5	17.6	20.3	18.2	9.7
BGE-VL-v1.5-mmeb	LlaVA-1.6-7B	8.2	18.6	19.5	19.7	17.3	2.0	14.0	9.9	12.3	12.6	12.4	6.6
BGE-VL-MLLM-S2	LlaVA-1.6-7B	7.6	17.3	18.3	13.6	15.7	1.7	14.0	9.8	12.6	10.5	12.2	5.3
VLM2Vec	Qwen2-VL-7B	21.6	38.5	32.1	39.0	35.5	11.1	31.9	24.8	25.5	31.3	26.7	19.6
GME	Qwen2-VL-7B	13.9	30.0	21.7	26.6	26.5	5.1	28.0	20.2	18.0	19.7	18.3	14.4
Ops-MM-E	Qwen2-VL-7B	11.3	26.8	23.7	23.3	24.0	4.8	24.9	18.0	17.1	16.6	17.0	13.1
Rzen-E	Qwen2-VL-7B	9.7	24.4	20.9	27.9	22.2	3.3	19.7	14.1	16.0	14.5	15.7	11.6

Table 8: The FG-CDE results for various models on the Text-to-Image (T2I) task of Flickr30k-CCD.

Model	Backbone	Text to Image											S&P
		Entity					Scene			Event			
		Type	Attr	Rel	Emo	Avg	Type	Attr	Avg	Event	Event	Avg	
CLIP-series embedding models													
CLIP	ViT-B(149M)	21.7	34.1	33.3	34.1	32.2	8.9	28.9	22.1	32.5	34.5	32.9	14.3
BGE-VL-base	ViT-B(149M)	24.3	33.5	32.2	31.9	31.9	7.7	27.5	20.7	34.6	35.5	34.8	14.9
BLIP-coco	ViT-B(149M)	17.9	28.6	28.7	31.5	27.2	5.0	23.5	17.2	24.1	27.5	24.8	13.5
BLIP-flickr30k	ViT-B(149M)	18.5	27.2	29.7	31.7	26.5	8.6	28.1	21.5	26.8	27.1	26.9	15.8
Align-base	EfficientNet+BERT	16.9	32.6	27.4	28.1	29.4	7.3	28.3	21.1	26.7	30.0	27.4	15.6
CLIP	ViT-L(307M)	18.7	34.8	31.0	29.9	31.7	10.6	30.7	23.8	28.3	31.2	28.9	15.5
BGE-VL-large	ViT-L(307M)	24.1	34.3	32.6	31.2	32.4	7.4	27.8	20.8	31.1	34.1	31.7	14.7
BLIP-coco	ViT-L(307M)	17.1	28.1	28.1	34.5	27.0	5.1	23.0	16.9	23.7	27.0	24.4	12.3
BLIP-flickr30k	ViT-L(307M)	15.9	25.4	26.6	33.8	24.8	4.8	23.6	17.2	24.3	27.3	24.9	12.7
SigLIP	SoViT(428M)	17.4	30.8	31.3	31.9	29.0	10.3	30.4	23.5	26.5	29.8	27.2	16.4
AltCLIP	ViT-H(632M)	17.1	33.4	28.3	29.1	30.2	9.1	29.2	22.3	26.5	30.7	27.4	14.6
MLLM-based embedding models													
GME	Qwen2-VL-2B	15.8	34.3	25.7	33.8	30.7	8.4	32.8	24.5	22.1	25.7	22.9	20.1
VLM2Vec	Qwen2-VL-2B	17.9	34.6	29.9	31.2	31.4	8.6	31.9	24.0	25.1	27.0	25.5	17.2
Ops-MM-E	Qwen2-VL-2B	14.4	31.5	27.1	29.7	28.4	7.7	29.9	22.3	24.0	25.4	24.3	15.5
UniME	Phi3.5-V-4.2B	15.7	30.0	25.4	27.2	27.2	6.7	26.9	20.0	22.4	24.6	22.9	14.1
VLM2Vec	LlaVA-1.6-7B	16.5	32.2	26.3	30.5	29.2	7.9	30.1	22.5	22.7	24.6	23.1	14.0
UniME	LlaVA-1.6-7B	14.9	30.9	23.4	26.9	27.5	5.9	26.8	19.7	19.8	23.2	20.5	13.8
UniME-OneVision	LlaVA-1.6-7B	11.0	28.3	21.9	24.0	24.8	5.4	24.7	18.1	17.9	20.7	18.5	12.5
BGE-VL-MLLM-S1	LlaVA-1.6-7B	18.2	30.9	29.4	33.0	29.1	4.3	24.0	17.3	24.7	28.9	25.6	11.6
BGE-VL-v1.5-mmeb	LlaVA-1.6-7B	11.8	26.9	22.8	25.4	24.2	4.8	24.0	17.4	19.7	20.7	19.9	12.8
BGE-VL-MLLM-S2	LlaVA-1.6-7B	13.4	28.1	23.3	24.9	25.2	5.4	24.6	18.1	20.8	24.3	21.5	12.3
VLM2Vec	Qwen2-VL-7B	16.4	32.1	25.7	28.6	28.9	7.8	28.5	21.4	19.8	23.8	20.7	14.2
GME	Qwen2-VL-7B	16.7	34.3	26.0	35.6	31.0	9.1	33.2	25.0	21.7	24.1	22.2	21.5
Ops-MM-E	Qwen2-VL-7B	13.2	29.9	26.3	25.1	26.7	5.7	27.3	19.9	21.0	24.4	21.7	14.8
Rzen-E	Qwen2-VL-7B	12.6	27.4	22.4	28.7	24.8	5.8	24.5	18.1	17.1	19.2	17.6	14.0

Table 9: The FG-CDE results for various models on the Image-to-Text (I2T) task of Flickr30k-CCD.

Task	Prompt
Contrastive Strategy Generation	Given a source image, the objective is to generate several contrastive strategies and corresponding information for contrastive images. Each contrastive image should preserve most of the original visual features while diverging in one key detail, thereby forming a contrastive pair with the source image. The contrastive aspects should be selected from the following aspects.\n {Definition of Taxonomy}\n Instructions:\n 1. Identify Key Details: Based on the image, identify 3-4 key details in the image and output their corresponding aspects. Ignore details that are difficult to obtain contrastive images and less important details in the given detailed descriptions. These key details should preferably span different dimensions and aspects. For example:\n Event Element: two Labradors are chasing the cat; Scene Attribute: The color of the lawn is green, hinting at summer.\n 2. Design Contrastive Strategies and Output Key Information: For each key detail, execute the following steps:\n - Output current contrastive detail.\n - Ouput aspect of current key detail from the above aspects.\n - Design Contrastive Strategy: Based on the contrastive detail, propose a strategy for generating the contrastive image. The strategy must perturb the source image by changing current contrastive detail, such as shifting Entity Type from Labrador to Husky, Event Element from dogs chasing cat to cat chasing dogs, Entity Attribute from summer green lawn to land covered with snow in winter, or Style and Presentation from normal to retro classic style. Note that the contrastive strategy should be realistic and achievable.\n - Output Image Editing Instruction: For each strategy, generate a clear, detailed editing instruction. An image editing tool will transform the original image into a contrastive image according to the editing instruction.\n {examples}
Caption Refinement	In the given images, there are both similarities and subtle differences. The text below provides a coarse-grained caption for the first image, which does not capture the fine-grained content, especially the differences from the other images. Your task is to generate a fine-grained caption for each image, following the style of caption below. \n Coarse-grained Caption of the First Image:\n {Caption}\n Instructions:\n 1. Identify the most significant differences between the image pairs. Limit a maximum of two key differences for each image pair. If one detail difference between two images is significant, while the other is minor, then the less noticeable difference can be ignored. Avoid noting minor discrepancies. Focus on the Entity (Type, Attribute, Relationship), Scene (Type, Attribute, Spatial Relationship), Event/Action (Category, Element, Attribute), Emotion and Mood, Style and Presentation. If the images depict sports such as surfing, skiing, or skateboarding, describe the specific posture.\n 2. Summarize all the key differences into key aspects of differences.\n 3. Based on the key aspects summarized in above step, identify the specific key features of each individual image.\n 4. For each image, following the style of the given coarse-grained caption, provide a fine-grained and fluent caption. The caption to ensure compliance with criteria below.\n Separating each caption with "" and adding a index (such as 1. or 2.) before each caption.\n {examples}\n The caption for image must meet the following criteria:\n (1) Consistency with Content: The caption for each image should reflect its content. While not every detail needs to be covered, it must accurately represent the essential elements of the image.\n (2) Highlighting Differences: The caption must emphasize the differences between this image and the others. At least one distinguishing feature should be included, ensuring the caption reflects a divergence from the other images.

Table 10: Prompt templates for the generation of contrastive strategies and refined captions.