Vision Projector: Improving Zero-Shot Composed Image Retrieval at Inference

Hoang-Bao Le, Allie Tran
Dublin City University
ADAPT Centre
Dublin, Ireland
bao.le2@mail.dcu.ie, allie.tran@dcu.ie

Binh T. Nguyen
Ho Chi Minh University of Science
Vietnam National University
Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn

Liting Zhou, Cathal Gurrin
Dublin City University
ADAPT Centre
Dublin, Ireland
{liting.zhou, cathal.gurrin}@dcu.ie

***Abstract*—Composed Image Retrieval (CIR) involves retrieving a target image based on a query composed of a reference image and a textual modification. Zero-Shot CIR extends this task by removing the need for labeled triplets during training. Most state-of-the-art (SOTA) methods share a common structure: a vision-language encoder followed by a matching module using Transformers or contrastive learning. Instead of increasing data or model complexity, we wonder that: Can we improve retrieval performance at inference time? To answer this, we propose the Vision Projector (VP)—a lightweight, plug-and-play module that enhances visual representations without retraining. Integrated directly into MagicLens, VP consistently improves performance across CIRR, FashionIQ, and CIRCO. Notably, it boosts MagicLens by 18% on CIRCO, despite not using its strongest variant. Code is available at: *https://github.com/baohl00/VisionProjector_ZSCIR*.**

***Index Terms*—composed image retrieval, zero-shot, vision projector.**

I. Introduction

Image Retrieval focuses on finding relevant content—such as images, text, or multimodal data—from large datasets. A specialized task, Composed Image Retrieval (CIR), retrieves target images based on a query formed by a reference image and a modifying text. This bi-modal query provides more precise intent, making CIR highly applicable to real-world systems like Google Lens or Google Photos.

However, CIR training requires triplet data—reference image, text query, and target image(s)—which is expensive and time-consuming to collect. To address this, Zero-Shot Composed Image Retrieval (ZS-CIR) [1] removes the need for human-labeled triplets, allowing for diverse, self-constructed training sets. Recent approaches like MagicLens [2], Pic2Word [3], and LaSCo [4] utilize large-scale web data or auto-generated datasets from sources like COCO or CC3M [5]. Most methods rely on Vision-Language Models (VLMs) such as CLIP [6] and BLIP [7], with some enhancing the queries using synthetic captions or LLMs [8]–[10].

Despite architectural differences, SOTA ZS-CIR models typically follow a two-stage pipeline: (1) encoding the image-text pair using a VLM, and (2) fusing the embeddings via a Transformer or MLP-based module. Yet, these pipelines often require model retraining to improve performance. In this work, we explore a different direction: *Can we enhance the visual features during inference to boost retrieval accuracy, without retraining the model on other datasets?*

To this end, we propose a set of lightweight Vision Projectors (VPs)—MLP-based modules applied to the encoded visual features. These VPs refine the representations by emphasizing task-relevant components. We also introduce an expansion layer to increase feature dimensionality, improving expressiveness. Our plug-and-play approach yields consistent gains across three standard CIR benchmarks—FashionIQ, CIRR, and CIRCO—surpassing original model baselines without altering their training process.

II. Related Work

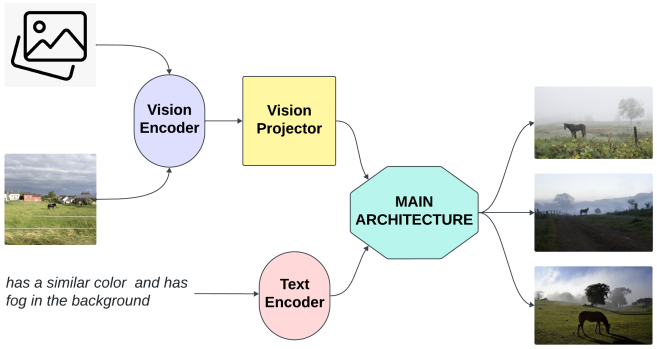

Fig. 1: Overview of combining Vision Projector in ZS-CIR. Normally, after the encoding stage, the encoded features are fused into the main architecture. However, with Vision Projector, we apply this before feeding them into the core architecture in order to enhance the feature of input images.

Zero-Shot Composed Image Retrieval. Composed Image Retrieval (CIR) is a multimodal task that retrieves a target image given a reference image and a textual modification. Datasets span fashion [11], [12], synthetic compositions [1], and natural scenes [4], [13], [14], all structured as triplets. CIR was first introduced by Vo *et al.* [1], using ResNet and LSTM encoders to learn a joint embedding space. Subsequent methods like CLIP4Cir [15] and LaSCo [4] adopted CLIP [6] and BLIP [7] for improved cross-modal alignment.

To avoid costly triplet annotations, Zero-Shot CIR (ZS-CIR) trains models without manual labels. Pic2Word [3] introduced the idea via contrastive learning on CC3M. Later works like SEARLE [16], i-SEARLE [14], and MagicLens [2] improved training via distillation, pseudo-token learning, or massive synthetic triplets. Others—e.g., LinCIR [10], LDRE [17], and FTI4CIR [9]—enhanced queries using LLM-generated captions or subject-attribute decomposition.

While most methods focus on data construction or training architectures, our approach improves performance by enhancing vision features at inference using a lightweight Vision Projector module—no retraining required.

Vision Projector. In the LLMs era, researchers have turned to improving their power by training LLMs with more parameters, more data and multi-GPUs. Liu *et al.* [18] also applied the Multilayer Perceptron (MLP) vision-language connector - a two-layer MLP that can improve LLaVA [19]'s multimodal capabilities in comparison to the linear projection. Especially, Yao *et al.* [20] proposed the Dense Connector, which is used as a plug-and-play vision language connector and provides the LLM with more visual cues. And recently, Cha *et al.* [21] has identified the crucial properties of visual projectors and introduced a locality-enhanced projector named Honeybee, which achieves better performance across the various MLLM benchmarks. Inheriting this idea, *Vision Projector* is designed to avoid training the model again, as well as to develop the visual understanding of the model.

In this paper, by using CLIP [6], we aim to accelerate visual factors and show the effect of expanding the embedding dimension on the result. In order to analyse the efficacy of *Vision Projector* (VP), we reproduce MagicLens and integrate it after the encoding step. Then, we compare the performance of our method among MagicLens's models and against TransAgg [8] and RTD+LinCIR [10], [22] in three datasets: FashionIQ [11], CIRR [13] and CIRCO [16].

III. Methodology

A. Overview

In this part, we illustrate the overview of MagicLens [2]. MagicLens has been introduced as a series of self-supervised image retrieval models that support open-ended instructions. The training data is a large crawled image dataset with about 36.7M triplets (reference image, text and target image). MagicLens outperforms the SOTA methods on the CIRCO dataset despite having a $50\times$ smaller model size. This efficiency is due to the parameter-sharing design in MagicLens and the training data advantage, including its large volume and high-quality paired images.

MagicLens uses CoCa [23] or CLIP [6] as the backbones for vision and language encoders, while other models such as SEARLE [16], i-SEARLE [14] or PLI [24] are generally applied on CLIP [6] or BLIP [7]. To enable deep modality integration, the authors designed multiple layers of self-attention followed by a single multi-head attention pooler. They also employ an empty text string for each image in the image pool not containing any relevant description. In the training period, a simple contrastive loss is used. However, one more thing is that the reference image itself is not removed from the ranking stage and this accidentally lowers the metric score because the textual query targets the other images sharing the same context with the reference one.

B. Vision Projector

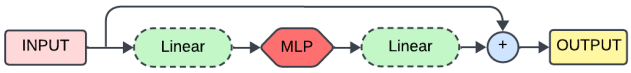

Fig. 2: General Vision Projector architecture. The input is the encoded and normalised features of images. Then the features go through the MLP architecture to enhance the important positions.

In this section, we describe **Vision Projector** (VP). Previous studies have shown that the embedding features [25], [26] from the pre-trained models such as CLIP often show an inconsistent relationship between their cluster. Therefore, Chen *et al.* [27] and Bao *et al.* [28] added Gaussian noise to enhance the representation of visual and textual features. Figure 1 shows how we apply VP into MagicLens. As it is designed to improve the features of the vision encoder, we plug it directly after this stage. The key point of using VP is that the parameters of MagicLens trained by Zhang et. al. [2] does not contain VP and we only utilise it in the inference stage. In figure 2, we illustrate the general design of VP, which is based on the Residual architecture. This specific architecture plays the most crucial role in our idea as it contains itself consistent relationships between existing features, and the MLP is used as a tool which highlights the most important features. Inspired by MLP and gating-MLP [29] (gMLP), we design them into new versions as Residual-N-MLP and gMLP-k-combine (gMLP-k-cb) that take advantage of their abilities.

Residual-N-MLP (Res-N-MLP) A MLP architecture consists of two fully-connected layers and a GeLU one . Here, we redesign a residual MLP in two ways: (1) By increasing the layers of MLP, we choose $N \in \{1, 2, 4\}$ and call them ResMLP, ResPhi (Res2MLP) and Res4MLP. There are two noticeable things that: with the lower embedding dimension of CLIP (512), the higher the number of layers, the better the model learns vision features. However, with CLIP$_{large}$, the embedding dimension is 768, and the results reach a peak if there are 2 MLP-layers. (2) The second way is to expand the input dimension by k times before feeding it into the MLP layer(s). In the experiment, we set $k \in \{4, 8, 12, 16\}$. It can be observed that both two base and large version of CLIP in MagicLens works well with $k = 8$ or $k = 12$. Overall, the ResPhi results provide more stability than other ResMLP-based vision projectors in all three datasets and the selected values of k.

gMLP$_k$-cb: gating-MLP (gMLP) has been introduced by Liu *et al.* [29], which is built out of basic MLP layers

with gating. gMLP shows comparable performance with DeiT [30] and Vision Transformer (ViT) [31] despite having fewer parameters. One notable point in gMLP is that instead of using `GELU` as activation function, we apply `Sigmoid` [32], [33] to reduce unnecessary patterns and highlight the more important ones, but keep the entire feature still coherent. Furthermore, for the gMLP architecture, we present three types of gMLP: (1) gMLP originally; (2) gMLP with k expansion rate (gMLP$_k$) and (2) gMLP combining with gMLP$_k$ (gMLP$_k$-cb). In gMLP$_k$, we increase the dimension of the input feature k times before going to the next stage. In gMLP$_k$-cb (Figure 3), we take the average of the original characteristic and gMLP$_k$. This approach ensures that the input keeps itself the main feature and still learns the gMLP features as well as the expansion gMLP ones.

```
def gmlp_combine(x, k):
    x_gmlp_original = gmlp_block(x)
    x_gmlp_k = gmlp_block(x, k)
    x_combine = (x_gmlp_original + x_gmlp_k)/2
    return x_combine

def gmlp_block(x, d_model, d_ffn, k = 1):
    shortcut = x
    if k > 1:
        d_ffn = d_ffn * k
    x = norm(x, axis="channel")
    x = proj(x, d_ffn_new, axis="channel")
    x = sigmoid(x) # or gelu(x)
    x = spatial_gating_unit(x)
    x = proj(x, d_model, axis="channel")
    return x + shortcut

def spatial_gating_unit(x):
    u, v = split(x, axis="channel")
    v = norm(v, axis="channel")
    n = get_dim(v, axis="spatial")
    v = proj(v, n, axis="spatial", init_bias=1)
    return u * v
```

Fig. 3: Python implementation of a gMLP$_k$-cb block with a spatial gating unit

IV. Experiment and Discussion

A. Experimental Setups

Dataset and Evaluation Metrics.

Following the ZS-CIR results of MagicLens [2], we evaluate our experiments using three benchmark datasets including FashionIQ [11], CIRR [13] and CIRCO [16]. We show the detailed statistic of used datasets on table I.

Dataset		# Query	# Index
FashionIQ [11]	Shirt	2,038	6,346
	Dress	2,017	3,817
	Toptee	1,961	5,373
CIRR [13]		4,148	2,316
CIRCO [16]		800	123,403

TABLE I: Dataset Statistics.

1) **FashionIQ** [11] contains ∼77.6k fashion images across three categories: Shirt, Dress, and Toptee, with over 30k triplets. The data is split into train, validation, and test sets in a 6:2:2 ratio.
2) **CIRR**[1] [13], built on NLVR^2 [34], includes 21.5k real-world images and 36k+ query-target pairs spanning nine relation types (e.g., Negation, Spatial Relations, Cardinality). Following prior work [2], [8], [22], we use Recall@K (R@K) for evaluation.
3) **CIRCO**[2] [16], based on COCO [35], features 120k+ images and 1,020 triplets, each with one reference image, a caption, and a target image. Queries have multiple ground-truth targets (avg. 4.53), and performance is measured using mean Average Precision at K (mAP@K).

Implementation details. All experiments are performed on one NVIDIA A100 GPU with Python 3.9 and Flax [36]. We use the visual and textual encoders of the CLIP ViT-B16 and ViT-L14 [6] from a JAX library named Scenic [37], and set image resolution of 224×224.

B. Baselines

Besides comparing with MagicLens [2], we include other state-of-the-art ZS-CIR models: TransAgg [8], RTD+LinCIR [10], [22], and MLLM-I2W [28]. TransAgg is a transformer-based model trained on 32k triplets, using both template-based and GPT-3.5-generated captions. LinCIR introduces self-masking projection to reduce reliance on triplet inputs, while RTD enhances text encoder generalization. MLLM-I2W further improves robustness via uncertainty modeling with Gaussian noise. We also compare with several CLIP-L-based methods: Pic2Word [3], i-SEARLE [14], CIReVL [38], and PLI [24].

C. Main Results

Comparison with SOTAs. Table II presents our results of the experiment on the evaluation data sets. By integrating the Vision Projector with MagicLens trained on the CLIP-L backbone, we achieve notable improvements across all benchmark datasets compared to the original MagicLens. On the CIRCO dataset, this configuration surpasses the previous state-of-the-art score achieved by CoCa-L-based MagicLens by over 1.2% and the original CLIP-L-based MagicLens by about 20%. For the FashionIQ and CIRR datasets, our model demonstrates consistent gains over MagicLens's performance, achieving higher metric scores than the baseline MagicLens. Overall, the average score improves by more than 3%, ranking third in the results table. Our model trails behind RTD+LinCIR and CoCa-L-based MagicLens by only 0.5% and 1.9%, respectively, while offering a more balanced improvement across all datasets. These results emphasize the effectiveness of the Vision Projector in enhancing retrieval performance, particularly on CIRCO.

[1]https://cirr.cecs.anu.edu.au/
[2]https://circo.micc.unifi.it

Method	Backbone	FashionIQ		CIRR		CIRCO		Average
		R@10	R@50	R@5	R@10	mAP@10	mAP@25	
Pic2Word [3]	CLIP-L	25.3	44.9	51.5	64.1	9.1	10.1	34.2
i-SEARLE [14]	CLIP-L	29.2	49.5	54.0	64.7	13.6	15.4	37.7
CIReVL [38]	CLIP-L	28.6	48.6	52.3	64.9	19.1	20.9	39.1
PLI [24]	CLIP-L	35.4	57.4	54.6	67.6	11.6	13.0	39.9
MLLM-I2W [28]	CLIP-L	30.3	50.1	57.9	70.2	-	-	-
TransAgg [8]	BLIP-B	34.4	55.1	**68.9**	**79.6**	30.9	32.2	50.2
RTD+LinCIR [22]	CLIP-G	**46.2**	**67.3**	67.5	78.3	22.3	24.5	51.0
MagicLens [2]	CoCa-L	38.0	58.2	67.0	77.9	35.4	38.1	**52.4**
	CLIP-L	30.7	52.5	61.7	74.4	30.8	33.4	47.3
+ Vision Projector	CLIP-L	32.1	53.1	64.9	76.9	**36.8**	**39.3**	50.5

TABLE II: Baseline comparison on three benchmarks. While we reproduce the results of TransAgg and MagicLens related to CLIP-L, the others are from the original papers. **Bold** numbers indicate the best results.

CLIP	Projector	FashionIQ			CIRR			CIRCO
		R@10	R@50	Avg.	R@1	R@5	R@50	mAP@5
B		26.38	48.57	37.47	31.3	61.52	92.05	25.46
	ResMLP	24.94	45.58	35.26	28.7	59.01	90.96	22.10
	ResPhi	26.51	47.94	37.23	31.42	60.72	91.95	24.37
	Res4MLP	26.59↑0.13	48.57	37.58↑0.11	31.52↑0.22	61.81↑0.29	91.98↓0.07	25.69↑0.23
	gMLP-GELU	17.65	36.60	27.12	20.72	50.07	86.94	14.84
	gMLP-σ	23.78	45.11	34.44	28.02	56.72	89.78	20.81
	gMLP$_k$	23.41	44.64	34.03	29.37	58.82	90.96	20.80
	gMLP$_k$-cb	25.95	46.93	36.44	29.28	59.25	90.68	22.86
L		30.82	52.06	41.44	33.28	63.83	*93.11*	30.12
	ResMLP	31.13	51.72	41.43	34.48	63.81	93.06	32.35
	ResPhi	31.69	*52.95*	*42.32*	35.11	*64.84*	93.06	34.06
	Res4MLP	31.26	52.31	41.79	33.35	63.95	**93.21**	31.69
	gMLP-GELU	28.00	49.17	38.59	30.84	59.88	90.82	29.57
	gMLP-σ	*31.95*	52.57	42.26	35.25	63.57	92.75	32.68
	gMLP$_k$	31.26	52.67	41.97	*35.46*	64.75	92.46	*34.51*
	gMLP$_k$-cb	**32.13**↑1.31	**53.07**↑1.01	**42.60**↑1.16	**35.74**↑2.46	**65.06**↑1.23	92.84↓0.27	**35.82**↑5.7

TABLE III: Best results of every Vision Projector combined with two MagicLens version. **Bold** and *Italic* numbers indicate the best and second values. The red and blue numbers represent the gap with the original model shared the same CLIP version.

Comparison between Vision Projectors. The best performance of each Vision Projector (VP) is shown in Table III. Combining the Vision Projector (VP) with MagicLens yields consistent improvements across both the Base (CLIP-B) and Large (CLIP-L) backbones. For CLIP-B, Residual MLP (ResMLP) architectures outperform gMLP variants, indicating their effectiveness in handling lower-dimensional features. For example, on CIRCO, the best numbers on gMLP-based belongs to gMLP$_k$-cb (mAP@5=22.86%) that are lower than MagicLens with Res4MLP (mAP@5=25.69%) and MagicLens without Vision Projector (mAP@5=25.46%). Increasing the number of layers in the MLP leads to higher performance, highlighting the benefit of deeper architectures in this setting. On the other hand, for CLIP-L-based MagicLens, ResPhi achieves superior results compared to ResMLP and Res4MLP, demonstrating its ability to better leverage the larger encoder dimensions. However, when comparing Residual MLP-based with gMLP-based, we observe that gMLP-based make a stronger impact on MagicLens's performance. For instance, on FashionIQ, the average score of gMLP$_k$-cb (42.60%) is 1.16% and 0.28% higher than none Vision Projector's (41.44%) and ResPhi's (42.32%).

Another key observation is the impact of activation functions. gMLP with GELU activation performs significantly worse than gMLP with Sigmoid activation, with performance gaps of approximately 6.0% and 3.0% for CLIP-B and CLIP-L, respectively. This underscores the importance of activation selection in optimising gMLP's performance. Furthermore, significant improvements are observed with gMLP, gMLP$_k$, and gMLP$_k$-cb for the CLIP-L backbone. Among these, gMLP$_k$-cb achieves the best results across almost all metrics, with a remarkable improvement on CIRCO (mAP@5=35.82%) compared to the original score of 30.12%.

V. Conclusion

We introduce the Vision Projector, a lightweight MLP-based module inspired by residual learning, to enhance visual embeddings in retrieval tasks. It significantly improves performance by refining visual features—especially with larger encoders—demonstrating strong scalability. Integrated directly into MagicLens without retraining, our approach outperforms it across FashionIQ, CIRR, and notably CIRCO. These results highlight Vision Projector as a practical, effective add-on. Future work will explore integration with other encoders like BLIP or CoCa and extend to new domains.

Acknowledgment

This publication has emanated from research supported in part by research grants from Science Foundation Ireland under grant numbers SFI/13/RC/2106_P2 and 18/CRT/6223, and co-funded by the European Regional Development Fund.

REFERENCES

[1] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval - an empirical odyssey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, "MagicLens: Self-supervised image retrieval with open-ended instructions," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 59 403–59 420. [Online]. Available: https://proceedings.mlr.press/v235/zhang24an.html

[3] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," *CVPR*, 2023.

[4] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski, "Data roaming and quality assessment for composed image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 2991–2999, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28081

[5] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565. [Online]. Available: https://aclanthology.org/P18-1238

[6] 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.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[8] Y. Liu, J. Yao, Y. Zhang, Y. Wang, and W. Xie, "Zero-shot composed text-image retrieval," *arXiv preprint arXiv:2306.07272*, 2023.

[9] H. Lin, H. Wen, X. Song, M. Liu, Y. Hu, and L. Nie, "Fine-grained textual inversion network for zero-shot composed image retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2024, pp. 240–250.

[10] G. Gu, S. Chun, W. Kim, , Y. Kang, and S. Yun, "Language-only training of zero-shot composed image retrieval," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[11] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback," *CVPR*, 2021.

[12] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1463–1471.

[13] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2125–2134.

[14] L. Agnolucci, A. Baldrati, M. Bertini, and A. Del Bimbo, "isearle: Improving textual inversion for zero-shot composed image retrieval," *arXiv preprint arXiv:2405.02951*, 2024.

[15] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Effective conditioned and composed image retrieval combining clip-based features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 21 466–21 474.

[16] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347.

[17] Z. Yang, D. Xue, S. Qian, W. Dong, and C. Xu, "Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 80–90.

[18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[20] H. Yao, W. Wu, T. Yang, Y. Song, M. Zhang, H. Feng, Y. Sun, Z. Li, W. Ouyang, and J. Wang, "Dense connector for mllms," *arXiv preprint arXiv:2405.13800*, 2024.

[21] J. Cha, W. Kang, J. Mun, and B. Roh, "Honeybee: Locality-enhanced projector for multimodal llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[22] J. Byun, S. Jeong, W. Kim, S. Chun, and T. Moon, "Reducing task discrepancy of text encoders for zero-shot composed image retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2406.09188

[23] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2205.01917

[24] J. Chen and H. Lai, "Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval," 2023. [Online]. Available: https://arxiv.org/abs/2311.07622

[25] S. Gu, C. Clark, and A. Kembhavi, "I can't believe there's no images! learning visual tasks using only language supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2672–2683.

[26] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.

[27] Y. Chen, Z. Zheng, W. Ji, L. Qu, and T.-S. Chua, "Composed image retrieval with text feedback via multi-grained uncertainty regularization," in *International Conference on Learning Representations (ICLR)*, 2024.

[28] T. Bao, C. Liu, D. Xu, Z. Zheng, and T. Xu, "MLLM-I2W: Harnessing multimodal large language model for zero-shot composed image retrieval," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Association for Computational Linguistics, 2025.

[29] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," *Advances in neural information processing systems*, vol. 34, pp. 9204–9215, 2021.

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[31] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] R. Csordás, K. Irie, and J. Schmidhuber, "Approximating two-layer feedforward networks for efficient transformers," 2023. [Online]. Available: https://arxiv.org/abs/2310.10837

[33] H. Nguyen, N. Ho, and A. Rinaldo, "Sigmoid gating is more sample efficient than softmax gating in mixture of experts," 2024. [Online]. Available: https://arxiv.org/abs/2405.13997

[34] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," 2019. [Online]. Available: https://arxiv.org/abs/1811.00491

[35] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[36] J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee, "Flax: A neural network library and ecosystem for JAX," 2024. [Online]. Available: http://github.com/google/flax

[37] M. Dehghani, A. Gritsenko, A. Arnab, M. Minderer, and Y. Tay, "Scenic: A jax library for computer vision research and beyond," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 393–21 398.

[38] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "Vision-by-language for training-free compositional image retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2310.09291