InternVL-VPR: Hierarchical Zero-Shot Visual Place Recognition with VLM-driven Re-Ranking

Zhi Hu[1], Liang Liao[2], and Weisi Lin[1]
[1]College of Computing & Data Science, Nanyang Technological University, Singapore
[2]Hangzhou Institute of Technology, Xidian University, China
E-mail: {zhi004, wslin}@ntu.edu.sg, liang.liao@xidian.edu.cn

***Abstract*—Vision language model (VLM)-aided visual place recognition (VPR) has recently emerged as a promising paradigm, where pre-trained vision encoders provide coarse retrieval and VLM reasoning refines candidate rankings. Despite recent progress, existing approaches employ giant trillion-parameter VLMs to deliver qualitative same-place judgments, without explicit reasoning, which may overlook informative visual cues. To address this, we propose a hierarchical two-stage zero-shot VPR framework that incorporates quantitative vision–language reasoning while leveraging a compact vision–language model. In the first stage, dense local features from vision foundation models are aggregated to form compact global descriptors for coarse retrieval. In the second stage, the VLM is prompted to generate discriminative local context descriptions together with importance scores that explicitly highlight location-identifiable cues. These descriptions enable quantitative re-ranking of candidates through weighted similarity computation. The framework remains fully zero-shot, requiring no task-specific fine-tuning and relying on a compact off-the-shelf 38B-parameter VLM. Experiments on the SF-XL Occlusion, SF-XL Night, and AmsterTime benchmarks demonstrate that our method achieves performance comparable to supervised approaches.**

I. Introduction

Visual Place Recognition (VPR) enables autonomous agents and augmented reality systems to re-localize by matching a query image to a database of geo-tagged images [1]–[6]. While recent supervised VPR methods have achieved impressive performance under controlled settings, their effectiveness deteriorates in out-of-distribution scenarios, such as severe appearance changes, occlusions, night-time illumination drop, and long-term temporal shifts [7]–[9]. Recently, vision language model (VLM)-aided VPR has emerged as a promising paradigm. It employs foundation vision models for coarse retrieval, followed by VLM reasoning to refine candidate rankings [10]. Despite exploiting rich pre-trained knowledge without task-specific fine-tuning, these approaches face some limitations: (*i*) they rely on giant trillion-parameter VLMs that are computationally expensive and impractical for edge-device deployment, and (*ii*) their re-ranking depends on direct qualitative same-place judgments, may lead to inconsistent rankings and fail to attend to critical visual cues.

To address these challenges, we propose a hierarchical two-stage zero-shot VPR framework that integrates quantitative vision–language reasoning with a compact VLM. In Stage 1, a pre-trained vision foundation model [11] extracts dense local descriptors from both query and database images, which are aggregated using a train-free VLAD [12] module into compact global descriptors for efficient coarse retrieval of the top-K nearest neighbors. In Stage 2, a 38B-parameter VLM [13] is prompted to produce textual descriptions of the most discriminative local regions, each paired with an attention-derived weight. These descriptions highlight stable landmarks, textures, and spatial arrangements while suppressing transient or low-texture areas. Each textual description is then embedded into a vector via an off-the-shelf embedding

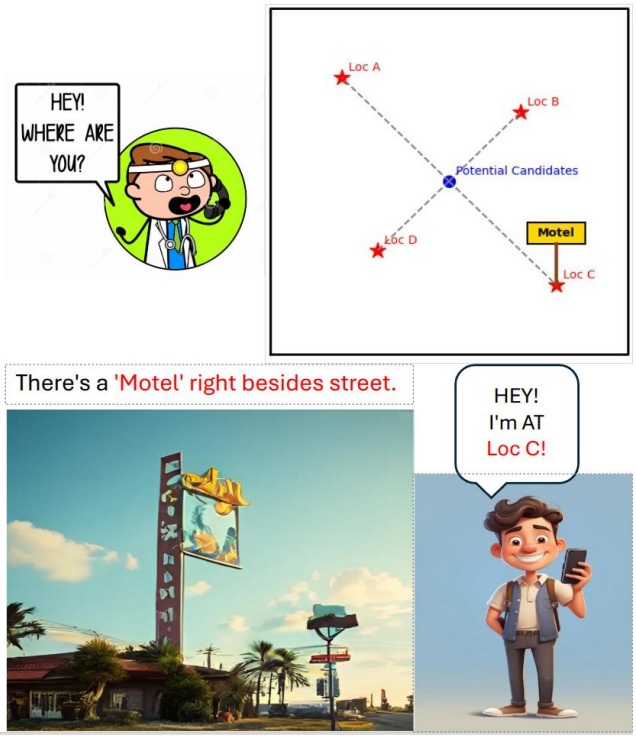

Fig. 1: Illustration of the proposed VPR pipeline. Given a query image from the agent, coarse retrieval shortlists candidate locations A–D. The refinement stage employs InternVL3-38B to generate textual descriptions for each candidate (e.g., identifying a "Motel" at Loc C). By comparing the weighted similarity, the system re-ranks the candidates and correctly selects Loc C.

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

Fig. 2: The proposed two-stage zero-shot VPR framework. Stage 1: The query image and the geo-tagged reference set are encoded into global descriptors using DINOv2 features aggregated by VLAD [4]. The top-K closest candidates are then retrieved based on descriptor similarity. Stage 2: A pretrained vision–language model (InternVL3-38B) generates discriminative textual descriptions for the query and each candidate image. These textual descriptions are transformed into embedding vectors through weighted aggregation and used for quantitative re-ranking of the K retrieved candidates, producing the final re-ranked result.

model [14]. The re-ranking is achieved through the aggregation of embeddings weighted by their corresponding importance scores. Experiments demonstrate that the proposed fully zero-shot and training-free framework achieves performance comparable to supervised counterparts on challenging benchmarks, while addressing the limitations of current LLM-aided VPR.

In summary, our contributions are threefold:

- We introduce a two-stage framework that integrates coarse retrieval with VLM-aided reasoning for hierarchical place recognition.
- We propose a VLM-based reasoning module that produces discriminative local context descriptions and corresponding attention scores, enabling a principled quantitative measure of match quality.
- Experiments on SF-XL Occlusion, SF-XL Night, and AmsterTime show that our framework achieves performance comparable to supervised counterparts while consistently surpassing zero-shot baselines.

II. Related Work

A. Conventional VPR

Conventional VPR methods primarily focused on designing robust global descriptors and aggregation schemes to address appearance and viewpoint variations. NetVLAD [15] pioneered a trainable VLAD layer on top of CNN backbones, significantly improving retrieval accuracy through metric learning. Subsequent work explored alternative feature extractors and aggregation strategies: MixVPR [1] combined a CNN backbone with an MLP-Mixer, outperforming NetVLAD; classification-framed methods such as CosPlace [8] and EigenPlace [16] leveraged dense place labels and orientation cues to improve discriminability. More recently, vision foundation models such as DINOv2 [11] have been adopted as off-the-shelf feature extractors. Extensions like SALAD [2] replaced the backbone with DINOv2 and applied optimal transport to model feature–cluster relationships, with clique mining further improving separation of similar scenes [3]. Despite the gains, these approaches often require task-specific fine-tuning, which limits their robustness in uncontrolled real-world scenarios.

B. Two-Stage Re-Ranking

Two-stage pipelines improve retrieval by re-ranking coarse results with fine-grained matching. Patched-NetVLAD [17] performs global retrieval followed by patch-level re-evaluation. Transformer-based architectures such as TransVPR and R2Former [18], [19] employ cross-attention mechanisms to align spatial features and reorder candidate lists more accurately. Although these approaches improve recall metrics, they incur substantial computational overhead and generally require training, making them less suitable for zero-shot deployment.

C. VLM-driven VPR

Large language models have recently been adopted in vision tasks [20]–[23], inspiring zero-shot VPR frameworks that use VLMs for re-ranking [10]. However, as discussed above, these approaches rely on narrative assessments and trillion-parameter models, limiting their reliability and practicality. In contrast, our framework addresses these issues by integrating train-free VLAD-based global retrieval with quantitative VLM-based reasoning, using a compact 38B-parameter VLM.

III. Methods

We propose a two-stage zero-shot VPR pipeline that first performs coarse image retrieval and then refines the results using a pre-trained VLM. In Stage 1, query and database images are encoded into a global descriptors [4]. The query descriptor is compared against database descriptors to retrieve a shortlist of top-K candidate images. In Stage 2, the pre-trained VLM [13] is prompted to generate context-aware textual descriptions for the query and each of the K candidates.

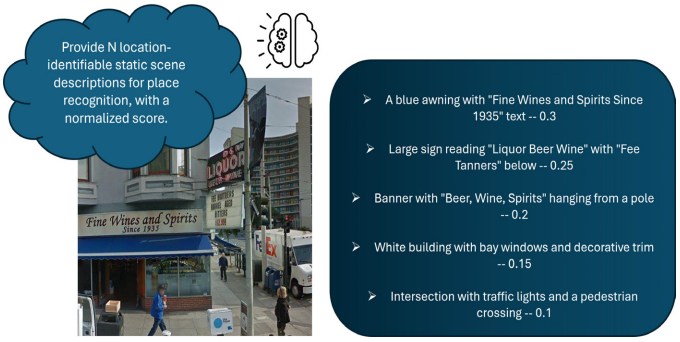

Fig. 3: Illustration of the discriminative local context description generation process. InternVL3-38B [13] is prompted to provide location-identifiable static scene descriptions, each associated with a normalized attention-derived importance score.

By comparing the weighted similarity, the candidates are re-ranked to produce the final match. Fig. 2 illustrates the overall workflow of our method, from coarse visual retrieval to fine-grained re-ranking via VLMs.

A. Coarse Retrieval Phase

We first build a visual codebook of S cluster centers $\{\mathbf{c}_s\}_{s=1}^{S}$ via k-means on a large set of DINOv2-G ViT [11] local descriptors. Given a query image I_q, we extract M local descriptors $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^{M}$. We compute soft-assignments to all centers using cosine-similarity weights:

$$a_{i,s} = \frac{\exp\big(\mathbf{x}_i^\top \mathbf{c}_s\big)}{\sum_{t=1}^{S} \exp\big(\mathbf{x}_i^\top \mathbf{c}_t\big)}, \tag{1}$$

where all vectors are L_2-normalized. We then accumulate weighted residuals:

$$\mathbf{v}_s = \sum_{i=1}^{M} a_{i,s}\,(\mathbf{x}_i - \mathbf{c}_s), \quad \mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_S] \in \mathbb{R}^{SD}. \tag{2}$$

Perform intra-normalization by

$$\mathbf{v}_s \leftarrow \frac{\mathbf{v}_s}{\|\mathbf{v}_s\|_2}. \tag{3}$$

concatenate $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_S]$, and then apply global normalization:

$$\tilde{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}. \tag{4}$$

Let $\tilde{\mathbf{v}}_q$ be the query descriptor and $\{\tilde{\mathbf{v}}_j\}$ precomputed for the database. We rank by

$$\mathrm{sim}(\tilde{\mathbf{v}}_q, \tilde{\mathbf{v}}_j) = \tilde{\mathbf{v}}_q^\top \tilde{\mathbf{v}}_j. \tag{5}$$

and select the top-K candidates for the re-ranking phase.

B. VLM-Driven Re-ranking

After coarse retrieval yields the top-K candidates $\{I_j\}_{j=1}^{K}$, we prompt InternVL3-38B [13] on both the query I_q and each candidate I_j to extract location-aware semantic cues. Specifically, as depicted in Fig. 3, we prompt the VLM with:

> "Provide N location-identifiable static scene descriptions for place recognition, each with a normalized score."

As a result we obtain:

$$T^q = \{(t_i^q, \alpha_i^q)\}_{i=1}^{N}, \quad T^j = \{(t_i^j, \beta_i^j)\}_{i=1}^{N},\ j = 1, \ldots, K, \tag{6}$$

where

$$\sum_{i=1}^{N} \alpha_i^q = 1, \quad \sum_{i=1}^{N} \beta_i^j = 1.$$

Here, t_i^q (resp. t_i^j) is the i-th descriptive phrase for the query (resp. candidate), and α_i^q, β_i^j are the normalized attention weights.

We then embed each phrase via a pretrained Sentence-BERT encoder $E(\cdot)$, normalizing to unit length:

$$\mathbf{e}_i^q = \frac{E(t_i^q)}{\|E(t_i^q)\|_2}, \quad \mathbf{e}_i^j = \frac{E(t_i^j)}{\|E(t_i^j)\|_2}. \tag{7}$$

To quantify how well I_j matches cue i, we compute:

$$s_i(I_q, I_j) = \max_{\ell=1,\ldots,N} \cos\big(\mathbf{e}_i^q, \mathbf{e}_\ell^j\big). \tag{8}$$

Picking the best-aligned candidate descriptor. We then fuse these cue scores using the query attention:

$$S(I_q, I_j) = \sum_{i=1}^{N} \alpha_i^q\ s_i(I_q, I_j). \tag{9}$$

Finally, we sort the K candidates by descending similarity score $S(I_q, I_j)$, retaining only those with

$$S(I_q, I_j)\ >\ \tau, \tag{10}$$

where τ is a predefined threshold to prevent erroneous re-ranking.

IV. Experiments

A. Datasets and Evaluation Metrics

We evaluate our approach on three challenging benchmarks:

- **SF-XL Occlusion** and **SF-XL Night**, exhibiting severe occlusion and illumination variations.
- **AmsterTime**, containing historical-to-modern scene matching with significant temporal appearance shifts.

Performance is evaluated using the $Recall@1$ and $Recall@5$ metrics, with a retrieval counted as correct if the matched image lies within 25 m of the ground-truth location for the SF-XL dataset.

B. Implementation Details

We employ the DINOv2-G ViT model [11] as our vision backbone to extract patch-token features. For re-ranking, we prompt InternVL3-38B [13] to produce three descriptions per image. Each descriptor is encoded with Sentence-BERT [14] to obtain normalized embeddings. To balance accuracy and efficiency, Stage 1 returns the top 100 visual candidates. All components are used off-the-shelf in a fully zero-shot pipeline without any task-specific fine-tuning.

TABLE I: Recall@1 and Recall@5 on three challenging VPR benchmarks. The best result is in **bold**, and ↑ indicates improvement over the zero-shot one-stage baseline (AnyLoc).

Category	Method	SF-XL Occ		SF-XL Night		Amstertime	
		R@1	R@5	R@1	R@5	R@1	R@5
Supervised	CosPlace [8]	30.3	44.7	23.6	32.8	49.6	68.5
	MixVPR [1]	30.3	38.2	19.5	30.5	40.9	59.5
	EigenPlaces [16]	32.9	52.6	23.6	34.5	49.9	69.8
	SALAD-CM [3]	**44.7**	**64.5**	43.8	57.1	**56.2**	**78.7**
Zero-Shot	AnyLoc [4]	15.8	27.6	41.0	55.4	45.1	66.3
InternVL-VPR	Ours	28.9 (13.1↑)	40.8 (13.2↑)	**48.9 (7.9↑)**	**62.4 (7.0↑)**	47.6 (2.5↑)	69.9 (3.6↑)

Fig. 4: Qualitative comparison of retrieval results on four challenging night-time queries. From left to right: the input **Query**, top-1 matches by **SALAD-CM** and **AnyLoc** (both false and marked in red), and our **InternVL-VPR** results (correct matches in green).

C. Compare with Supervised Counterparts

Table I compares the proposed InternVL-VPR with recent supervised and zero-shot visual place recognition methods across three challenging benchmarks: SF-XL Occlusion, SF-XL Night, and AmsterTime. Our method consistently outperforms the zero-shot baseline AnyLoc [4], with Recall@1 improvements of 13.1%, 7.9%, and 2.5% on SF-XL Occlusion, SF-XL Night, and AmsterTime, respectively. These gains highlight the effectiveness of quantitative vision–language reasoning in refining retrieval beyond coarse global embeddings.

Compared with supervised methods, InternVL-VPR achieves comparable performance despite requiring no task-specific fine-tuning. On SF-XL Occlusion, our approach achieves Recall@1 close to CosPlace and MixVPR, while remaining entirely zero-shot. On AmsterTime, InternVL-VPR performs on par with EigenPlaces. Notably, on the SF-XL Night benchmark, our framework surpasses the state-of-the-art supervised model SALAD-CM [3], achieving Recall@1 of 48.9 versus 43.8. This result demonstrates the strong generalization capability of the proposed method under severe illumination changes, a scenario where supervised pipelines tend to struggle.

D. Ablation Study

1) Effect of Number of VLM Descriptions: Table II evaluates the effect of varying the number of VLM-generated discriminative descriptors (N) per image on the SF-XL Occlusion and SF-XL Night datasets. Results indicate that employing three descriptors per image consistently yields optimal Recall@1 and Recall@5 across both datasets. While increasing N from one to three descriptors enhances performance—indicating that a single descriptor is insufficient in complex scenarios—further increasing N to five does not imply further improvement. Instead, adding additional descriptors may introduce non-discriminative or redundant contextual information, adversely affecting the re-ranking accuracy. Hence, three descriptors per image represent an optimal balance between capturing essential visual context and avoiding performance degradation.

TABLE II: Ablation analysis with varying numbers of VLM-generated discriminative descriptors (N) per image. Optimal performance is highlighted in bold.

N	SF-XL Occ		SF-XL Night	
	R@1	R@5	R@1	R@5
1	23.7	35.5	46.1	60.5
3	**28.9**	**40.8**	**48.9**	**62.4**
5	**28.9**	39.5	46.8	61.2

E. Qualitative Analysis

1) Retrieval in Challenging Settings: As depicted in Fig. 4, the proposed method leverages VLMs to extract distinct semantic cues—such as neon sign text (“Motel”, “Pizza”,

Fig. 5: Illustration of the quantitative re-ranking process. A query image (leftmost) and retrieved candidates (three columns to the right) are passed through a VLM to generate discriminative textual descriptions, each with normalized attention scores. Candidates are re-ranked by computing similarity between the query's description vector and each candidate's description vector. The candidate most closely matching the query's key descriptors (highlighted in green) is identified as the final match.

"Liquors") and unique architectural motifs—that remain robust under challenging conditions like low illumination and viewpoint changes. Unlike supervised counterparts such as SALAD-CM [3] and the zero-shot baseline AnyLoc [4], which frequently misidentify visually similar façades or repetitive textures, our approach generates place-identifiable textual descriptions that emphasize stable and discriminative cues. By re-ranking candidates based on the similarity of these descriptions, our method effectively disambiguates hard negative samples (e.g., buildings with similar layouts but different signage) and recovers correct matches overlooked by prior methods.

2) Re-ranking Process: Fig. 5 illustrates the core mechanism of our quantitative re-ranking stage. Given a query image (left), the VLM is prompted to generate structured textual descriptions of its most discriminative local regions, each accompanied by normalized attention weights that reflect their relative importance. These descriptions typically capture stable and location-specific cues, such as distinctive signage, architectural layouts, or environmental context.

The final ranking is obtained by aggregating these weighted similarities across all descriptors, thereby transforming narrative semantic information into quantitative metrics. This allows our framework to recover the correct candidate (green box) even when visually similar distractors (red boxes) share large portions of background structure, effectively disambiguating challenging cases where conventional supervised or zero-shot approaches fail.

V. Conclusion

We introduced InternVL-VPR, a fully zero-shot, two-stage hierarchical framework that combines DINOv2+VLAD coarse retrieval with a 38B-parameter InternVL3 model for re-ranking. Unlike prior LLM-aided methods that depend on trillion-parameter models and direct narrative judgments, our approach performs quantitative vision–language reasoning with a compact VLM, achieving accuracy comparable to supervised methods on challenging benchmarks without any task-specific fine-tuning.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (62202349), and the Fundamental Research Funds for the Central Universities (ZYTS25036).

REFERENCES

[1] A. Ali-Bey, B. Chaib-Draa, and P. Giguére, "Mixvpr: Feature mixing for visual place recognition," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2997–3006.

[2] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 658–17 668.

[3] S. Izquierdo and J. Civera, "Close, but not there: Boosting geographic distance sensitivity in visual place recognition," in *The European Conference on Computer Vision (ECCV)*, Oct. 2024.

[4] N. Keetha, A. Mishra, J. Karhade, *et al.*, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, 2024.

[5] I. Tzachor, B. Lerner, M. Levy, *et al.*, "EffoVPR: Effective foundation model utilization for visual place recognition," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[6] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[7] B. Yildiz, S. Khademi, R. M. Siebes, and J. Van Gemert, "Amstertime: A visual place recognition benchmark dataset for severe domain shift," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 2749–2755.

[8] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4868–4878.

[9] C. Sakaridis, D. Dai, and L. Van Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7373–7382.

[10] Z. Lyu, J. Zhang, M. Lu, Y. Li, and C. Feng, "Tell me where you are: Multimodal llms meet place recognition," *arXiv preprint arXiv:2406.17520*, 2024.

[11] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, ISSN: 2835-8856.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311, 2010.

[13] J. Zhu, W. Wang, Z. Chen, *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Nov. 2019.

[15] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.

[16] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 046–11 056.

[17] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 136–14 147.

[18] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 638–13 647.

[19] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R^2 former: Unified retrieval and reranking transformer for place recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 370–19 380.

[20] H. Wu, Z. Zhang, E. Zhang, *et al.*, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *ICLR*, 2024.

[21] Z. Zhang, H. Wu, E. Zhang, G. Zhai, and W. Lin, "Q-bench$^+$+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 404–10 418, 2024.

[22] H. Wu, Z. Zhang, E. Zhang, *et al.*, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 490–25 500.

[23] C. Li, H. Wu, Z. Zhang, *et al.*, "Q-refine: A perceptual quality refiner for ai-generated image," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.