PanoRec: Spatially-Structured Sequence Modeling for Multi-Granularity Panoramic Retrieval

Zidong Cao[1], Ding Zhou[2], Wenyao Gao[2], Lutao Jiang[1], and Hui Xiong[1,3⋆]
caozidong1996@gmail.com, deonzhou@tencent.com, tankygao@tencent.com, jianglutao98@gmail.com, xionghui@ust.hk

[1] AI Thrust, HKUST (Guangzhou), China
[2] WeChat, Tencent, China
[3] Department of CSE, HKUST, Hong Kong SAR, China

Abstract. Panoramic images capture holistic environments, yet retrieving them using fine-grained, localized textual descriptions remains a fundamental challenge. Through empirical analysis, we reveal that modern Vision-Language Models (VLMs) suffer from severe *semantic dilution* when processing panoramic inputs. By compressing an information-dense panorama into a single global embedding, VLMs inevitably submerge local details within vast backgrounds, restricting retrieval to coarse scene-level matching. To overcome this bottleneck, we propose **PanoRec**, a multi-granularity panoramic retrieval framework built on a *spatially-structured sequence modeling* paradigm. Specifically, PanoRec serializes distortion-free cubemap faces and a downsampled global panorama alongside spatial anchor tokens into a unified sequence. This enables the efficient extraction of decoupled local and global representations within a single forward pass. To effectively supervise this multi-granularity feature space, we formulate a *joint spatial InfoNCE* objective. For local matching, we adopt a MaxSim routing strategy that dynamically aligns each query with its most relevant cubemap face. Crucially, this strategy not only suppresses background noise during inference but also inherently introduces a powerful hard spatial negative mining mechanism during training. Extensive experiments demonstrate that PanoRec achieves impressive performance across multiple scenarios, effectively unifying holistic scene-level retrieval with fine-grained spatial discrimination.

Keywords: Panoramic Image · Cross-Modal Retrieval · Sequence Modeling · Multi-Granularity

1 Introduction

Panoramic images capture holistic 360°×180° environments, offering immersive visual context that is crucial for applications such as virtual reality (VR), autonomous navigation, and scene understanding. Meanwhile, the rapid evolution

⋆ Corresponding Author.

of large-scale Vision-Language Models (VLMs) [2, 25, 35] has revolutionized 2D visual reasoning. However, directly applying these powerful foundation models to panoramas is fundamentally hindered by the severe geometric distortions inherent in the standard Equirectangular Projection (ERP) format. While early works mitigated these distortions using specialized spherical operators [19, 28], such task-specific architectural intrusions irreversibly disrupt the powerful 2D visual priors acquired during massive-scale pre-training [3]. Consequently, projection-based methods that transform the spherical data into distortion-free perspective planes, such as cubemaps (CPs) [33], have emerged as the prevailing paradigm. By natively aligning 360° data with standard 2D inputs, these methods enable the direct leverage of VLMs without requiring any architectural alterations.

To leverage general-purpose VLMs for panoramic retrieval, a straightforward approach is to process the entire 360° image and compress it into a single global embedding. To systematically investigate the feasibility of this adaptation, we conduct an empirical pilot study evaluating a state-of-the-art (SOTA) VLM [16] across diverse retrieval granularities. Specifically, we evaluate the model on both aligned-granularity tasks (*e.g.*, global text to global panorama, local text to an isolated cubemap face) and cross-granularity tasks (*e.g.*, local text to a global panorama). Our analysis reveals a stark contrast: while the general VLM performs well when granularities are aligned, its performance degrades significantly during cross-granularity matching. We term this critical bottleneck *semantic dilution*: crucial localized visual details are inevitably submerged within the vast background context of the surrounding environment. This finding demonstrates that compressing an information-dense panorama into a single holistic embedding causes severe background interference. Consequently, it motivates our core design: explicitly preserving the spatial independence of local views to prevent semantic dilution, while simultaneously retaining the holistic context.

To this end, we propose **PanoRec**, a multi-granularity panoramic retrieval framework built upon a *spatially-structured sequence modeling* paradigm. Rather than compressing the information-dense panorama into a single holistic embedding, PanoRec serializes distortion-free cubemap faces and a downsampled global panorama into a unified input sequence. Crucially, each visual input is seamlessly interleaved with a learnable spatial anchor token (*e.g.*, `<front>`, `<global>`). During the forward pass, this spatially-structured serialization allows the base VLM to perform holistic cross-view reasoning. More importantly, instead of relying on complex multi-branch networks, PanoRec elegantly extracts a multi-vector representation, comprising decoupled view-specific local embeddings and a holistic global embedding, directly from the hidden states of these anchor tokens within a single forward pass. This early-fusion approach fully utilizes the native interleaved sequence capabilities of modern VLMs, achieving fine-grained spatial feature extraction without requiring any task-specific architectural modifications.

To effectively optimize this decoupled representation space, we formulate a joint spatial InfoNCE objective. A naive matching might aggregate the extracted view-specific embeddings via mean-pooling, but this inevitably entangles target visual features with irrelevant background context. Instead, we adopt a MaxSim

routing strategy for local feature matching. This mechanism dynamically computes the similarity score by aligning the local query with the single most relevant cubemap face, suppressing background noise during inference. Moreover, applying this MaxSim operation within the in-batch contrastive loss introduces a powerful hard spatial negative mining mechanism. During training, the query is contrasted not only against its ground-truth view but also against the highest-scoring cubemap faces from non-target panoramas. This effectively forces the model to acquire fine-grained spatial discrimination.

In summary, our main contributions are threefold. **(i)** First, we identify the critical bottleneck of *semantic dilution* when applying general-purpose VLMs to panoramic retrieval and propose PanoRec. This framework serializes distortion-free cubemaps and a downsampled global panorama into a unified spatially-structured sequence, enabling the extraction of multi-granularity representations in a single forward pass. **(ii)** Second, we formulate a joint spatial InfoNCE objective equipped with a dynamic MaxSim routing strategy. For local feature matching, this mechanism introduces a powerful hard spatial negative mining process to avoid shortcut learning. **(iii)** Finally, extensive experiments demonstrate that PanoRec achieves impressive performance across diverse retrieval granularities.

2 Related Work

2.1 Panoramic Image Representations

Panoramic images in the standard ERP format suffer from severe geometric distortions [1, 18]. Early efforts mitigated this via specialized distortion-aware operators [6, 15, 28, 29] or spherical attention mechanisms [17, 19, 39]. For instance, PanoSwin [19] introduces pitch attention and shifted window schemes for distortions and boundary discontinuities. Despite their success in specific tasks, these approaches necessitate task-specific architectural modifications. Such structural intrusions can compromise the integrity of large-scale Vision-Language Models (VLMs) [2, 16, 25, 35], disrupting the powerful 2D visual priors acquired during massive-scale pre-training [3]. Alternatively, projection-based methods transform the sphere into distortion-free perspective planes like cubemaps (CPs) [33, 34] and tangent patches (TPs) [8, 38], aligning with standard 2D VLM inputs. However, traditional projection methods often treat these planes as an unordered "Bag-of-Views" with parallel processing [8]. While recent methods [13, 14, 20] explicitly model geometric adjacency to enforce cross-view consistency, they ultimately yield *a single global embedding at the output level*. The single-vector paradigm inherently sacrifices spatial decoupling and hinders fine-grained retrieval. *To overcome this bottleneck, PanoRec serializes distortion-free cubemaps and a global panorama into a unified spatially-structured sequence, supporting both holistic scene comprehension and fine-grained localized retrieval.*

2.2 Panoramic Cross-Modal Retrieval

Despite the advanced spatial reasoning capabilities of modern VLMs [2, 25], applying them to panoramic retrieval exposes a unique representation bottleneck.

Compressing the information-dense panorama into a single global embedding sacrifices spatial decoupling, causing local details to be submerged within vast backgrounds. To mitigate this semantic dilution, the general 2D vision community has shifted toward dense patch matching [10, 37]. In the panorama domain, Dense360 [40] extracts dense token embeddings directly from the ERP grid. However, these dense paradigms introduce significant computational bottlenecks. Storing thousands of dense vectors per panorama induces intractable index bloat and expensive MaxSim latency [10]. Furthermore, our multi-granularity task fundamentally differs from cross-view geo-localization [27, 41]. While the latter matches overhead satellite imagery to ground-level panoramas to solve extreme viewpoint transformations, our focus remains on localizing fine-grained semantics within a complicated 360° environment. *Drawing inspiration from multi-vector paradigms [12], PanoRec introduces a multi-vector representation comprising six decoupled view-specific local embeddings and a holistic global embedding.*

2.3 Contrastive Representation Learning

Standard cross-modal retrieval heavily relies on global contrastive learning objectives, typically the InfoNCE loss [23, 25], to align visual and textual representations. To improve discriminative power, sophisticated hard negative mining strategies [5, 9] have been widely explored in traditional 2D vision-language tasks. However, directly applying standard batch-wise contrastive learning to panoramic domains frequently leads to severe shortcut learning [26]. Because inter-panorama negative samples often originate from entirely different environments, models easily distinguish them by exploiting trivial cues, such as global color palettes and ambient lighting, instead of learning geometric structures. Consequently, the learned representations lack the fine-grained structural awareness required to localize view-specific semantics. *To overcome it, PanoRec formulates a joint spatial InfoNCE objective. By contrasting the local query against the most confusing view-specific faces across the batch, we force the model to look beyond global cues and acquire fine-grained spatial discrimination.*

3 Empirical Analysis of Semantic Dilution

Overview. While recent VLMs demonstrate remarkable zero-shot reasoning on standard perspective images, their capacity for fine-grained panoramic retrieval has not been systematically examined. To explicitly identify the representational bottlenecks in this domain, we conduct a series of zero-shot pilot studies.
Protocol. We employ the pre-trained Qwen3-VL-Embedding models (2B and 8B) and evaluate them on real-world indoor (Matterport3D [4]) and outdoor (CVACT [21]) datasets. All experiments report the Recall@1 (R@1) metric.

3.1 Granularity Asymmetry on Complete Panoramas

We first investigate how query granularity affects retrieval when the candidate pool consists of panoramas in the ERP format. We evaluate two types of textual

Table 1: Retrieval performance (R@1 (%)) on complete panoramas on both Matterport3D and CVACT datasets. We employ global and local captions for evaluation.

Dataset	Model	Global Caption	Local Caption (Six Cubemap Faces) Front	Right	Back	Left	Up	Down	Avg.
Matterport3D [4]	2B	**35.9**	8.7	4.8	2.6	4.4	0.4	2.0	3.8
	8B	**62.1**	13.6	7.1	3.5	6.8	1.2	4.8	6.2
CVACT [21]	2B	**11.7**	4.4	3.5	2.6	3.2	0.2	0.2	2.4
	8B	**41.6**	11.9	7.4	5.2	7.6	0.6	0.5	5.5

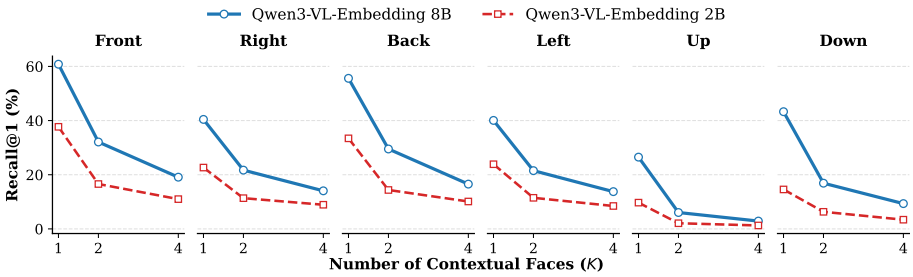

Fig. 1: Retrieval performance (R@1 (%)) under varying background context sizes ($K \in \{1, 2, 4\}$). For K=1 (*Oracle View*), the VLM processes the single target face. For $K \in \{2, 4\}$, non-target background faces from the exact same panorama are progressively stitched to construct 2D image grids. We evaluate on the Matterport3D.

queries: *Global Captions* that describe the holistic scene, and *Local Captions* that characterize fine-grained structural details of a specific cubemap face.

As shown in Tab. 1, taking the 8B model as an example, the zero-shot VLM achieves an R@1 of 62.1% (indoor) and 41.6% (outdoor) when queried with global captions, indicating that VLMs inherently possess strong scene-level matching capabilities. However, when switching to local captions that target the exact same panoramas but describe only a localized view, the performance drops to 6.2% and 5.5%, respectively. We hypothesize that this significant *granularity asymmetry* stems from a fundamental representational bottleneck: when a highly specific local query is matched against a compressed, information-dense panoramic embedding, the localized target signals might be submerged by the overwhelming background content.

3.2 Impact of Background Interference

To verify this hypothesis, we investigate how fine-grained semantic alignment behaves when background pixels are progressively introduced. We project the ERP images into distortion-free cubemap faces and fix the query to *local captions*. We then vary the number of cubemap faces fed into the model, denoted as

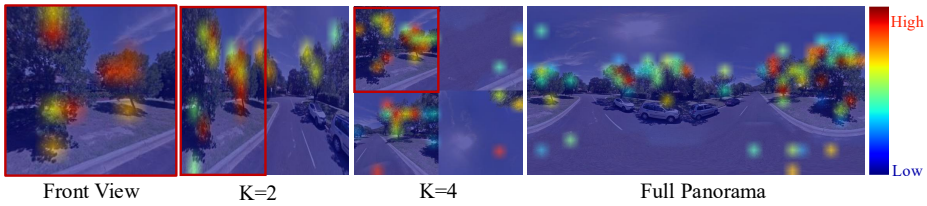

Fig. 2: Cross-Modal Similarity under Varying Context Sizes. When queried with a localized token (*e.g.*, "trees"), the similarity is concentrated on the target region for the front target view. However, as K increases, the similarity disperses to irrelevant regions like the sky.

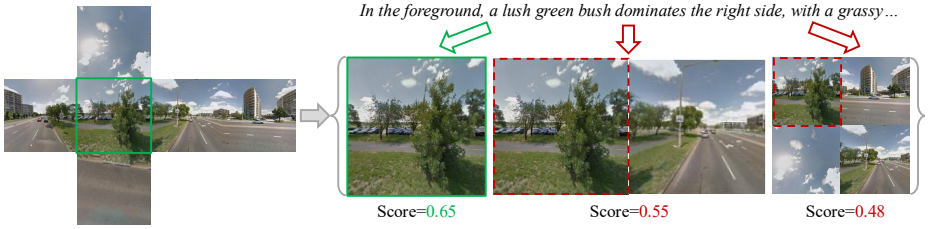

Fig. 3: Degradation of Local-to-Global Semantic Alignment. Given a fixed local caption describing the target view, its cosine similarity with the image embedding degrades as more non-target faces are stitched into the input.

$K \in \{1, 2, 4\}$. Here, K=1 serves as the *Oracle View*, feeding the VLM the single target cubemap face. For $K \in \{2, 4\}$, we introduce deterministically sampled non-target faces from the same panorama to construct 2D image grids.

As depicted in Fig. 1, the VLM achieves a robust R@1 under the K=1 oracle setting. This confirms that zero-shot VLMs are fully capable of understanding fine-grained local semantics of panoramas when presented in a single cubemap face. However, as the number of cubemap faces K increases, the performance exhibits a monotonic degradation. This effectively validates our hypothesis: background content from information-rich panoramas acts as substantial visual interference. When an entire panorama is forced into a single global embedding, dense background pixels dominate the representation, suppressing the localized targets and causing irreversible *semantic dilution*.

3.3 Qualitative Evidence of Semantic Dilution

We provide qualitative results to further verify this phenomenon. First, we compute the cross-modal similarity between a specific textual token (*i.e.*, "trees") and each cubemap face grid to examine how the model distributes spatial focus. As shown in Fig. 2, for the oracle face ($K = 1$), the similarity is sharply concentrated on the relevant objects. Conversely, as additional background faces are introduced ($K = 2$ and $K = 4$), the similarity maps become increasingly

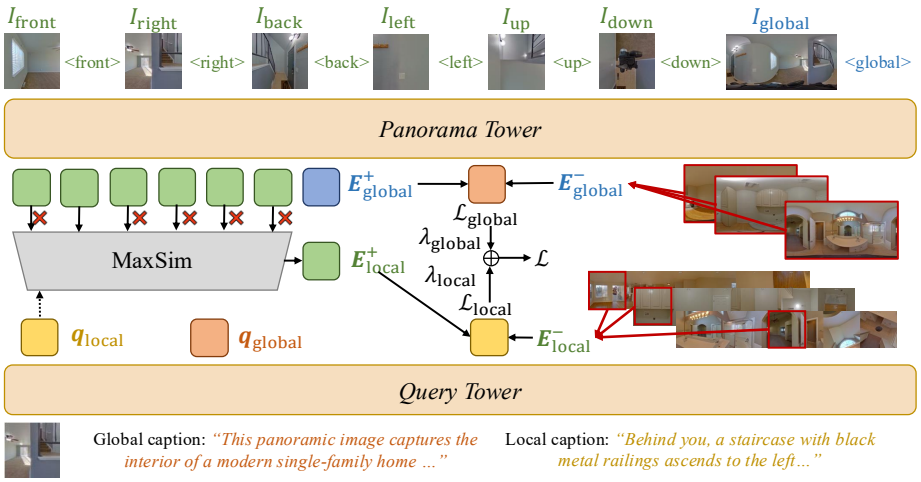

Fig. 4: The overall architecture of PanoRec. The *Query Tower* encodes incoming search queries (*e.g.*, narrow-FoV image crops, global and local captions) to generate query embeddings. The *Panorama Tower* extracts decoupled multi-granularity embeddings from a unified sequence. To optimize this decoupled representation space, a joint spatial InfoNCE objective utilizes dynamic MaxSim routing for local matching, suppressing background noise and enabling hard spatial negative mining.

fragmented, with high activations dispersing into irrelevant background pixels. Furthermore, Fig. 3 demonstrates that given a fixed local caption, the global cosine similarity between the text and the aggregated image embedding degrades significantly as more background context is stitched into the input. This empirically illustrates how the localized target signal is suppressed when compressed into a single embedding, providing direct visual evidence of the semantic dilution.
Summary. The quantitative and qualitative analyses consistently demonstrate that forcing a VLM to compress an entire panorama into *a single embedding* dilutes view-specific fine-grained semantics due to dominant background content. These findings motivate our design to bypass the representational bottleneck: rather than compressing a 360° scene into one embedding, we should preserve view-specific representations while simultaneously aggregating global context.

4 Methodology

Overview. To enable multi-granularity panoramic retrieval and mitigate semantic dilution, we introduce *PanoRec*, a dual-encoder framework initialized from Qwen3-VL-Embedding [16]. As illustrated in Fig. 4, the retrieval pipeline consists of two parallel branches. First, a *Query Tower* encodes incoming search queries (*e.g.*, text descriptions or narrow-FoV image crops) into dense query embeddings, denoted as $\boldsymbol{q}_{\text{global}}$ or $\boldsymbol{q}_{\text{local}}$ depending on their granularity. Meanwhile, the *Panorama Tower* processes the panorama via a spatially-structured

sequence modeling paradigm (Sec. 4.1). Specifically, it serializes distortion-free cubemap faces and a downsampled global panorama into a unified sequence $\mathcal{S}$, where each visual input is followed by a spatial anchor token (*e.g.*, `<front>`, `<global>`). This formulation efficiently extracts a multi-vector representation, comprising six view-specific local embeddings $\boldsymbol{E}_{\text{local}} \in \mathbb{R}^{6\times D}$ and one global embedding $\boldsymbol{E}_{\text{global}} \in \mathbb{R}^{D}$, within a single forward pass. Finally, to optimize this representation space, we formulate a joint spatial InfoNCE objective composed of $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$ (Sec. 4.2). For local feature matching, we utilize a MaxSim routing strategy to compute the similarity score S_{local} by aligning $\boldsymbol{q}_{\text{local}}$ against the most relevant cubemap face in $\boldsymbol{E}_{\text{local}}$. This effectively suppresses background noise and enforces fine-grained spatial discrimination.

4.1 Spatially-Structured Sequence Modeling

The primary challenge in multi-granularity panoramic retrieval is extracting fine-grained local semantics without suffering from spherical distortions or the semantic dilution caused by compressing the entire panorama into a single global embedding. To address this, we introduce a spatially-structured sequence modeling paradigm that leverages the interleaved context capabilities of VLMs.

Cubemap Projection and Contextual Preservation. Directly processing the raw ERP image inevitably introduces spherical distortions. Therefore, we first project the ERP image into six independent, distortion-free cubemap faces, denoted as $\mathcal{F} = \{I_{\text{front}}, I_{\text{right}}, I_{\text{back}}, I_{\text{left}}, I_{\text{up}}, I_{\text{down}}\}$. While these faces capture view-specific local structures, discarding the original ERP format sacrifices holistic structural cues. To preserve this holistic awareness, we complement the cubemap faces with a downsampled version of the original ERP image, denoted as I_{global}, to balance performance and computational costs. Retaining this global visual context proves beneficial for both global and local retrieval.

Sequence Modeling. Existing methods typically process multiple panoramic projections using either computationally expensive parallel forward passes or complex attention mechanisms. Instead, we adopt an early-fusion serialization strategy. We explicitly expand the VLM's vocabulary with a set of learnable spatial anchor tokens: $\mathcal{T}_{\text{local}} = \{$`<front>`, `<right>`, `<back>`, `<left>`, `<up>`, `<down>`$\}$ and a global anchor token `<global>`. We then construct a unified, interleaved input sequence $\mathcal{S}$ by appending the corresponding anchor token immediately after each visual input:

$$\mathcal{S} = \big[\, I_{\text{front}}, \texttt{<front>}, \ldots, I_{\text{down}}, \texttt{<down>}, I_{\text{global}}, \texttt{<global>} \,\big]. \tag{1}$$

This spatially-structured serialization strictly enforces geometric order, ensuring that the VLM interprets the panorama not as an unordered "bag-of-views", but as a coherent scan of the surrounding environment.

Single-Pass Multi-Vector Extraction. During the forward pass, the VLM processes the entire unified sequence $\mathcal{S}$ simultaneously. To obtain the target representations, we directly extract the final hidden states corresponding to the exact positions of the six spatial anchor tokens and the global anchor token. This

efficiently yields a set of L_2-normalized embeddings, comprising six view-specific local embeddings $\boldsymbol{E}_{\text{local}} \in \mathbb{R}^{6\times D}$ from $\mathcal{T}_{\text{local}}$ and one global embedding $\boldsymbol{E}_{\text{global}} \in \mathbb{R}^D$ from `<global>`. Consequently, PanoRec extracts a highly expressive multi-granularity representation within a single forward pass. Importantly, this early-fusion paradigm fully leverages the interleaved sequence modeling capabilities of VLMs. By avoiding task-specific architectural modifications, PanoRec effectively preserves the powerful visual priors of the pre-trained foundation models.

4.2 Multi-Granularity Matching and Optimization

In this section, we introduce a matching and optimization strategy to supervise the extracted multi-vector representations and enable multi-granularity retrieval.
MaxSim Routing for Local Matching. Given a local query embedding $\boldsymbol{q}_{\text{local}}$, evaluating its relevance to a panorama requires matching it against the six view-specific embeddings $\boldsymbol{E}_{\text{local}}$. A straightforward approach is to aggregate these six vectors into a single panoramic representation via mean-pooling or attention mechanisms [32]. However, since a local query typically describes only a specific region of the environment, such forced fusion inevitably entangles the target visual features with irrelevant background context, leading to semantic dilution.

To address this, we adopt a MaxSim routing strategy. The similarity score S_{local} is computed by taking the maximum cosine similarity across all six faces:

$$S_{\text{local}}(\boldsymbol{q}_{\text{local}}, \boldsymbol{E}_{\text{local}}) = \max_{1\le i\le 6} \left(\frac{\boldsymbol{q}_{\text{local}} \cdot \boldsymbol{E}_{\text{local},i}}{\|\boldsymbol{q}_{\text{local}}\| \|\boldsymbol{E}_{\text{local},i}\|} \right). \tag{2}$$

This formulation acts as a dynamic routing mechanism. It automatically aligns the query with the most relevant cubemap face while naturally discarding visual noise from the remaining non-target faces.
Joint Spatial InfoNCE Objective. Building upon this similarity metric, we optimize the network using a joint spatial InfoNCE objective. The overall training loss is composed of two components: $\mathcal{L} = \lambda_{\text{global}}\mathcal{L}_{\text{global}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}}$. Here, $\mathcal{L}_{\text{global}}$ aligns the global text with the holistic embedding $\boldsymbol{E}_{\text{global}}$ using standard cosine similarity, while $\mathcal{L}_{\text{local}}$ aligns the local query with $\boldsymbol{E}_{\text{local}}$ using the MaxSim score S_{local}. Applying the MaxSim operation within the in-batch InfoNCE loss inherently introduces a *hard spatial negative mining* mechanism. During training, a local query is contrasted not only against its ground-truth face but also against the highest-scoring (*i.e.*, most confusing) face from every other panorama in the batch. This forces the model to learn fine-grained spatial discrimination.

5 Experiments

5.1 Experimental Setup

Datasets and Annotations. To evaluate multi-granularity panoramic retrieval across domains, we utilize two established panoramic benchmarks: ZInD [7] for indoor scenes and CVACT [21] for outdoor cityscapes. To further assess the

Table 2: Main results on ZInD and CVACT datasets. We compare PanoRec against the ERP baseline across backbones and training paradigms (LoRA, Full-FT).

Train	Method	Global Caption			Local Caption			Local Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ZInD Dataset										
LoRA	Baseline-2B	60.61	86.17	91.52	7.63	20.32	27.73	11.93	28.53	36.10
	Ours-2B	**68.96**	**90.97**	**95.21**	**61.98**	**83.15**	**88.68**	**88.88**	**95.78**	**97.24**
	Baseline-8B	68.60	90.90	95.08	11.50	28.18	36.86	15.36	35.49	44.09
	Ours-8B	**76.83**	**94.78**	**97.30**	**74.80**	**90.87**	**94.33**	**92.31**	**96.87**	**97.87**
Full	Baseline-2B	67.57	90.94	95.37	13.10	32.36	42.39	14.07	34.06	43.19
	Ours-2B	**75.91**	**94.56**	**97.26**	**69.76**	**88.61**	**92.87**	**78.43**	**87.12**	**89.92**
	Baseline-8B	71.49	92.86	96.37	17.33	39.55	49.96	16.16	38.55	48.24
	Ours-8B	**79.68**	**95.97**	**98.10**	**76.55**	**92.19**	**95.35**	**82.63**	**89.90**	**92.22**
CVACT Dataset										
LoRA	Baseline-2B	53.70	79.11	86.35	12.27	27.14	35.22	21.87	38.94	46.09
	Ours-2B	**56.40**	**81.26**	**88.14**	**55.03**	**76.21**	**83.35**	**97.17**	**99.33**	**99.64**
	Baseline-8B	59.98	83.78	89.98	15.77	33.43	42.13	30.55	50.90	58.36
	Ours-8B	**62.10**	**85.92**	**91.16**	**69.37**	**87.41**	**92.06**	**94.83**	**97.67**	**98.41**
Full	Baseline-2B	56.84	82.83	89.72	15.54	34.07	43.78	25.52	45.85	54.17
	Ours-2B	**61.47**	**85.55**	**91.31**	**66.93**	**86.48**	**91.50**	**93.26**	**95.58**	**96.43**
	Baseline-8B	57.79	83.60	90.26	16.37	36.05	46.16	27.66	48.52	56.76
	Ours-8B	**61.91**	**85.94**	**91.87**	**71.47**	**89.00**	**93.07**	**93.00**	**95.43**	**96.26**

zero-shot capabilities, we additionally evaluate the ZInD-trained model on the Matterport3D [4] dataset. Since the original datasets lack the fine-grained textual descriptions required for our fine-grained panoramic retrieval task, we annotate each panorama across these datasets with a tailored set of seven texts: one holistic scene-level caption and six fine-grained view-specific captions.

Evaluation Metrics. We assess PanoRec across three distinct cross-granularity and aligned-granularity tasks: Global Text-to-Panorama retrieval, Local Text-to-Global Panorama retrieval, and Local Image-to-Global Panorama retrieval. We report standard retrieval metrics Recall@K (R@1, R@5, and R@10).

Implementation Details. PanoRec is built upon the Qwen3-VL-Embedding backbone [16]. To evaluate the adaptation capabilities of our framework, we implement two distinct training paradigms: parameter-efficient fine-tuning and full-parameter fine-tuning. In both paradigms, the visual encoder remains strictly frozen to preserve the pre-trained 2D visual priors. In the parameter-efficient setup, we apply Low-Rank Adaptation (LoRA) [11] to the attention projection layers (Q, K, V, O) of the language model. In the full-parameter setup, all weights of the language model are updated. Regardless of the chosen paradigm,

Table 3: Zero-shot performance comparison on the Matterport3D dataset. We compare PanoRec against general-purpose VLMs and direct training baselines.

Method	Global Caption			Local Caption			Local Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Qwen3-VL-2B	35.93	69.58	80.61	3.80	11.70	17.18	9.39	26.78	35.93
Direct Train-2B	54.57	86.42	92.99	12.38	34.97	46.73	19.39	52.07	64.46
PanoRec-2B	**59.42**	**89.16**	**94.91**	**72.66**	**91.00**	**94.51**	**90.45**	**97.07**	**98.36**
Qwen3-VL-8B	62.14	91.60	96.00	6.96	21.00	29.09	12.50	33.88	44.60
Direct Train-8B	62.18	91.09	95.96	17.13	44.44	57.12	22.55	57.96	70.48
PanoRec-8B	**68.16**	**94.00**	**97.52**	**82.78**	**95.83**	**97.63**	**91.95**	**97.73**	**98.87**

the seven newly introduced spatial anchor tokens are configured as fully trainable parameters to explicitly learn view-specific structural representations. For the visual inputs, each cubemap face is resized to 280×280 pixels, yielding 100 visual tokens per face, while the downsampled global panorama is resized to 560×280 pixels. The framework is optimized using FlashAttention-2, bf16 mixed precision, and DeepSpeed ZeRO-2 distributed training across eight H20 GPUs.

5.2 Main Results and Analysis

Overall Performance. In Tab. 2, PanoRec achieves SOTA performance across all tasks, consistently outperforming both the zero-shot VLMs and the fine-tuned ERP baselines. Notably, PanoRec exhibits a substantial performance improvement in fine-grained tasks. In the ZInD dataset under the LoRA setting, PanoRec-8B achieves a remarkable 74.80% R@1 for Local Caption retrieval, an improvement over the 11.50% achieved by the ERP baseline. This consistent superiority across diverse backbones and training paradigms validates the robustness of our spatially-structured sequence modeling. Moreover, while PanoRec is designed to enhance local discrimination, it also yields significant gains in Global Caption retrieval. For instance, on CVACT with Full-FT, PanoRec-8B improves the Global R@1 from 57.79% to 61.91%. This suggests that the decoupled local details provide complementary cues that enrich the holistic representation.

Generalization. In Tab. 3, PanoRec demonstrates impressive zero-shot generalization and scaling properties. When evaluated on the unseen Matterport3D dataset, PanoRec-8B achieves 68.16% Global R@1 and 82.78% Local R@1, significantly surpassing the Direct Train-8B baseline (62.18% and 17.13%, respectively).

Generalization across VLM Backbones. We apply PanoRec to diverse VLM families and scales (SmolVLM-256M to InternVL3.5-8B). As shown in Tab. 4, PanoRec consistently outperforms the ERP baseline across all backbones, especially on fine-grained tasks (*e.g.*, InternVL3.5-8B Local Caption R@1 from 5.23% to 56.62%). This confirms that our sequence modeling broadly unlocks the multi-granularity potential of general-purpose VLMs.

Table 4: Results of utilizing more VLMs on the ZInD dataset. We compare PanoRec against the ERP baseline across backbones with LoRA fine-tuning.

Method	Global Caption			Local Caption			Local Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SmolVLM-256M [22]	29.64	57.68	68.79	2.49	8.00	12.03	3.72	10.24	14.26
PanoRec	**37.06**	**65.61**	**75.78**	**29.52**	**52.17**	**61.74**	**86.58**	**93.34**	**95.37**
SmolVLM-500M [22]	28.99	56.84	68.05	2.35	7.83	12.00	4.18	11.13	15.38
PanoRec	**40.86**	**68.90**	**78.36**	**31.59**	**54.08**	**63.30**	**86.40**	**93.24**	**95.11**
Qwen3.5-0.8B [24]	38.36	66.64	76.40	3.63	10.92	16.00	6.51	16.17	21.62
PanoRec	**51.52**	**79.16**	**86.64**	**45.89**	**69.80**	**77.61**	**89.99**	**95.40**	**96.71**
InternVL3.5-1B [36]	35.44	63.85	74.09	3.63	10.82	15.76	5.57	14.62	19.85
PanoRec	**50.27**	**77.81**	**85.63**	**43.18**	**67.02**	**75.35**	**86.60**	**93.40**	**95.07**
InternVL3.5-2B [36]	39.31	67.62	77.36	4.15	12.22	17.70	6.85	17.84	23.88
PanoRec	**57.21**	**83.54**	**89.64**	**51.25**	**74.02**	**81.22**	**82.40**	**90.41**	**92.84**
InternVL3.5-8B [36]	43.78	72.37	81.14	5.23	15.06	21.32	7.73	20.17	26.81
PanoRec	**61.46**	**85.86**	**91.59**	**56.62**	**78.38**	**84.76**	**88.85**	**94.42**	**95.87**
Gemma-3-4B [30]	42.44	71.18	80.04	4.68	13.92	20.16	6.97	19.42	26.10
PanoRec	**59.34**	**84.85**	**90.58**	**52.94**	**75.17**	**82.01**	**89.82**	**95.74**	**97.00**
Kimi-VL-A3B [31]	51.71	79.84	87.33	7.26	19.58	26.98	10.58	25.98	33.13
PanoRec	**60.28**	**85.49**	**91.60**	**56.58**	**78.52**	**84.88**	**85.29**	**92.64**	**94.71**

5.3 Ablation Studies

Input and Architecture Variants. We investigate the contribution of each component in PanoRec by comparing four architectural variants. **(i) ERP Baseline**: The model processes a single high-resolution ERP image (1120×560) and produces a one-vector global representation. **(ii) Single-Vector Cubemap**: The model takes six cubemap faces as input but aggregates them into a single holistic embedding at the output level. **(iii) Multi-Vector Cubemap**: The model adopts our spatially-structured sequence with six view-specific tokens but omits the downsampled global panorama. **(iv) Full PanoRec**: Our complete multi-granularity framework as described in Sec. 4.

Effectiveness of Spatially-Structured Representation. In Tab. 5, the transition from single-vector paradigms (i, ii) to multi-vector paradigms (iii, iv) yields a significant performance gain in local retrieval. While variant (ii) mitigates geometric distortion via cubemap projections, its R@1 for Local Caption remains at 18.95%, which is comparable to the ERP baseline (17.33%). This confirms our core hypothesis: the primary bottleneck in panoramic retrieval is not merely distortion, but *semantic dilution* caused by single-vector compression. By disentangling local views into independent embeddings, variant (iii) improves the Local R@1 to 74.78%. Comparing variants (iii) and (iv) highlights the vital role

Table 5: Ablation study on **input and representation variants**.

Variant	Global Caption			Local Caption			Local Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
(i)	71.49	92.86	96.37	17.33	39.55	49.96	16.16	38.55	48.24
(ii)	65.86	90.23	94.57	18.95	42.24	53.03	24.23	49.44	59.64
(iii)	71.91	93.12	96.42	74.78	91.43	94.83	82.46	90.48	92.79
(iv) (Ours)	**79.68**	**95.97**	**98.10**	**76.55**	**92.19**	**95.35**	**82.63**	**89.90**	**92.22**

Table 6: Ablation on the interactions **during scoring.**

Interaction	Global Caption			Local Caption			Local Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Mean	77.38	95.23	97.76	16.25	36.53	46.80	32.75	58.04	67.66
Attention	77.68	95.27	97.83	23.29	49.21	61.31	49.17	77.87	86.30
MaxSim	**79.68**	**95.97**	**98.10**	**76.55**	**92.19**	**95.35**	**82.63**	**89.90**	**92.22**

of the downsampled global panorama. The inclusion of the global context in our full model (iv) not only boosts Global R@1 from 71.91% to 79.68% but also provides non-trivial improvements in local tasks, such as increasing Local Image R@1 from 82.46% to 82.63%.

Effectiveness of MaxSim Mechanism. To verify that preserving spatial independence is crucial for fine-grained retrieval, we compare our MaxSim routing strategy against two common feature aggregation schemes: *Mean Pooling* and *Attention Pooling*. In these variants, the six view-specific embeddings are collapsed into a single vector before computing the contrastive loss or retrieval score. As reported in Tab. 6, both Mean and Attention pooling suffer from a catastrophic drop in local retrieval performance. Specifically, Mean pooling only achieves a dismal 16.25% R@1 for Local Caption, suggesting that averaging disparate views essentially "washes out" discriminative local details with irrelevant background noise. While Attention pooling improves the result to 23.29%, it still falls significantly short of PanoRec. In contrast, our MaxSim strategy yields a massive improvement, outperforming Attention pooling by 53.26% in Local Caption R@1. This provides empirical evidence that forced feature fusion is fundamentally sub-optimal. By dynamically routing each query to its most relevant face, MaxSim effectively circumvents semantic dilution and ensures that fine-grained structural information remains distinct and discriminative.

Impact of Batch Size on Contrastive Learning. In Fig. 5(a), a larger batch size steadily improves Global and Local Caption retrieval, as the InfoNCE objective benefits from richer in-batch spatial negatives, with gains saturating beyond a batch size of 16. Local Image retrieval exhibits a mild inverse trend rather than a uniform gain.

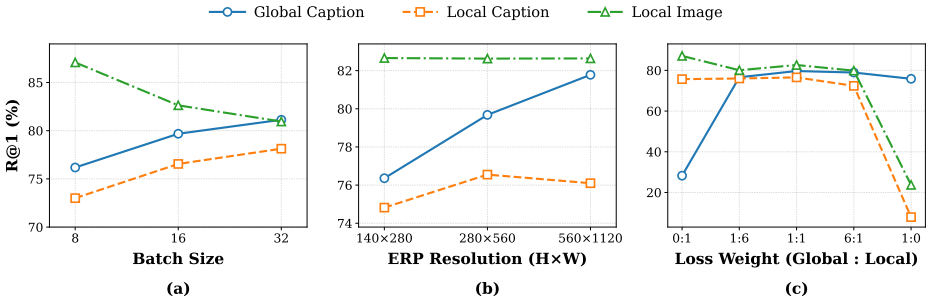

Fig. 5: Ablation Studies on **(a)** Batch size; **(b)** The resolution of downsampled global panorama. **(c)** The loss weights of global and local losses.

Table 7: Soft assignment results. We fine-tune the 8B backbone with LoRA on the ZInD dataset and report the R@1 score.

Method	Global Caption	Local Caption	Local Image
MaxSim (Default in PanoRec)	76.83	74.80	92.31
(i) Top-k avg, k=2	75.69	48.03	78.68
(i) Top-k avg, k=4	74.30	21.38	52.14
(ii) Adaptive-threshold top-k (δ=0.03)	76.65	74.60	91.57
(ii) Adaptive-threshold top-k (δ=0.05)	76.71	74.70	91.86
(iii) Dyn-τ ($\tau_{\min}$=0.01, $\tau_{\max}$=0.10, γ=50)	**77.07**	**75.29**	93.16
(iii) Dyn-τ ($\tau_{\min}$=0.01, $\tau_{\max}$=0.20, γ=50)	76.54	75.08	92.58
(iii) Dyn-τ ($\tau_{\min}$=0.03, $\tau_{\max}$=0.10, γ=50)	76.81	74.96	**93.57**
(iii) Dyn-τ ($\tau_{\min}$=0.01, $\tau_{\max}$=0.10, γ=30)	76.94	74.89	91.68

Influence of Global Panorama Resolution. The downsampled panorama is designed to provide a holistic context without redundant pixels. In Fig. 5(b), increasing the resolution from 140×280 to 560×1120 results in a steady improvement in the global retrieval. In contrast, Local Caption peaks at the 280×560 resolution and slightly declines. This suggests that excessive resolution in the global panorama may introduce redundant information that interferes with the precise view-specific details provided by the cubemap faces.

Analysis of Joint Loss Weights. We investigate the trade-off between global and local objectives by varying their loss weights (Global:Local). In Fig. 5(c), training with local loss alone (0:1) drops in global retrieval. Conversely, relying solely on the global loss (1:0) influences the model to localize specific views. Optimal performance across granularities is achieved with balanced ratios such as 1:1 or 1:6, showing that global comprehension and local discrimination are complementary tasks that benefit from simultaneous supervision.

Soft Assignment for Local Matching. Beyond the single most similar face, other faces may also provide useful evidence. We thus study soft assignments by training and testing several variants. Let $s_i = \cos(\boldsymbol{q}_{\text{local}}, \boldsymbol{E}_{\text{local},i})$ be the similarity between the local query and the i-th face. All strategies can be written as $S_{\text{local}}(\boldsymbol{q}_{\text{local}}, \boldsymbol{E}_{\text{local}}) = \sum_{i=1}^{6} w_i s_i$, differing only in the weights w_i, where MaxSim is the one-hot case ($w_i = 1$ for the most similar face only). We evaluate three soft assignments in Tab. 7: **(i)** *top-k averaging* uniformly averages the top-k faces; **(ii)** *adaptive-threshold aggregation* averages faces whose scores lie within δ of the best; and **(iii)** *dynamic-temperature softmax* uses $w_i = \exp(s_i/\tau)/\sum_j \exp(s_j/\tau)$ with a per-query temperature $\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min})\,\sigma(-\gamma((s_{(1)} - s_{(2)}) - m_0))$ determined by the top-1/top-2 ($s_{(1)}/s_{(2)}$) margin (m_0=0.05). The results show that top-k averaging fails on local tasks (Local Caption R@1 drops to 48.03%/21.38% for k=2/4), adaptive-threshold aggregation is on par with MaxSim, and dynamic-temperature softmax slightly improves over MaxSim (75.29% *vs.* 74.80%).

6 Conclusion

In this paper, we identify and address the critical bottleneck of *semantic dilution* in panoramic cross-modal retrieval. We reveal that conventional single-vector representations fail to capture the dense, multi-granularity information inherent in 360° environments. To overcome this, we propose PanoRec, a novel framework that leverages a spatially-structured sequence modeling paradigm to extract decoupled multi-granularity embeddings within a single forward pass. By integrating distortion-free cubemap faces with a global panorama, PanoRec effectively unifies fine-grained local discrimination with holistic scene comprehension. Furthermore, we introduce a joint spatial InfoNCE objective equipped with a dynamic MaxSim routing strategy, which inherently enables hard spatial negative mining and suppresses background noise. Extensive experiments on indoor and outdoor benchmarks, including zero-shot evaluations, demonstrate that PanoRec significantly outperforms existing panoramic and general-purpose VLM baselines. Our work provides an efficient solution for panoramic understanding, paving the way for more precise and context-aware immersive panoramic retrieval systems.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, in part by the guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (2024ZDZX1007). This work was supported by the project "Cross-Domain and Multimodal Generative Foundation Model for WeChat Live-Streaming Recommendation".

References

1. Ai, H., Cao, Z., Wang, L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision: H. ai et al. International Journal of Computer Vision **133**(8), 4973–5012 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Cai, Z., Yeh, C.F., Xu, H., Liu, Z., Meyer, G., Lei, X., Zhao, C., Li, S.W., Chandra, V., Shi, Y.: Depthlm: Metric depth from vision language models. arXiv preprint arXiv:2509.25413 (2025)
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 2017 International Conference on 3D Vision (3DV). pp. 667–676. IEEE Computer Society (2017)
5. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. Advances in neural information processing systems **33**, 8765–8775 (2020)
6. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European conference on computer vision (ECCV). pp. 518–533 (2018)
7. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2133–2143 (2021)
8. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12426–12434 (2020)
9. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
10. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449 (2024)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
12. Humeau, S., Shuster, K., Lachaux, M.A., Weston, J.: Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969 (2019)
13. Jung, D., Choi, J., Lee, Y., Manocha, D.: Rpg360: Robust 360 depth estimation with perspective foundation models and graph optimization. arXiv preprint arXiv:2509.23991 (2025)
14. Kalischek, N., Oechsle, M., Manhardt, F., Henzler, P., Schindler, K., Tombari, F.: Cubediff: Repurposing diffusion-based image models for panorama generation. In: The Thirteenth International Conference on Learning Representations (2025)
15. Khasanova, R., Frossard, P.: Geometry aware convolutional filters for omnidirectional images representation. In: International conference on machine learning. pp. 3351–3359. PMLR (2019)
16. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)

17. Li, X., Wu, T., Qi, Z., Wang, G., Shan, Y., Li, X.: Sgat4pass: Spherical geometry-aware transformer for panoramic semantic segmentation. arXiv preprint arXiv:2306.03403 (2023)
18. Lin, X., Ge, X., Zhang, D., Wan, Z., Wang, X., Li, X., Jiang, W., Du, B., Tao, D., Yang, M.H., et al.: One flight over the gap: A survey from perspective to panoramic vision. arXiv preprint arXiv:2509.04444 (2025)
19. Ling, Z., Xing, Z., Zhou, X., Cao, M., Zhou, G.: Panoswin: a pano-style swin transformer for panorama understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17755–17764 (2023)
20. Liu, A., Li, Z., Chen, Z., Li, N., Xu, Y., Plummer, B.A.: Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance. In: European Conference on Computer Vision. pp. 146–164. Springer (2024)
21. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5624–5633 (2019)
22. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)
23. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
24. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
27. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems **32** (2019)
28. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360 imagery. Advances in neural information processing systems **30** (2017)
29. Su, Y.C., Grauman, K.: Kernel transformer networks for compact spherical convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9442–9451 (2019)
30. Team, G.: Gemma 3 technical report. CoRR **abs/2503.19786** (2025). https://doi.org/10.48550/ARXIV.2503.19786, https://doi.org/10.48550/arXiv.2503.19786
31. Team, K.: Kimi-vl technical report. CoRR **abs/2504.07491** (2025). https://doi.org/10.48550/ARXIV.2504.07491, https://doi.org/10.48550/arXiv.2504.07491
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 462–471 (2020)
34. Wang, F.E., Yeh, Y.H., Tsai, Y.H., Chiu, W.C., Sun, M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5448–5460 (2022)

35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
37. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
38. Yun, H., Lee, S., Kim, G.: Panoramic vision transformer for saliency detection in 360 videos. In: European Conference on Computer Vision. pp. 422–439. Springer (2022)
39. Yun, I., Shin, C., Lee, H., Lee, H.J., Rhee, C.E.: Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6101–6112 (2023)
40. Zhou, Y., Zhang, T., Zhang, D., Ji, S., Li, X., Qi, L.: Dense360: Dense understanding from omnidirectional panoramas. arXiv preprint arXiv:2506.14471 (2025)
41. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)