

Region-based Cluster Discrimination for Visual Representation Learning

Yin Xie^{†1} Kaicheng Yang^{†1} Xiang An^{†1} Kun Wu¹ Yongle Zhao¹ Weimo Deng¹
Zimin Ran² Yumeng Wang¹ Ziyong Feng¹ Roy Miles³ Ismail Elezi³ Jiankang Deng^{4*}

¹ DeepGlint ² University of Technology Sydney

³ Huawei London Research Center ⁴ Imperial College London

{yinxie,kaichengyang,xiangan}@deepglint.com, jiankangdeng@gmail.com

Abstract

Learning visual representations is foundational for a broad spectrum of downstream tasks. Although recent vision-language contrastive models, such as CLIP and SigLIP, have achieved impressive zero-shot performance via large-scale vision-language alignment, their reliance on global representations constrains their effectiveness for dense prediction tasks, such as grounding, OCR, and segmentation. To address this gap, we introduce Region-Aware Cluster Discrimination (RICE), a novel method that enhances region-level visual and OCR capabilities. We first construct a billion-scale candidate region dataset and propose a Region Transformer layer to extract rich regional semantics. We further design a unified region cluster discrimination loss that jointly supports object and OCR learning within a single classification framework, enabling efficient and scalable distributed training on large-scale data. Extensive experiments show that RICE consistently outperforms previous methods on tasks, including segmentation, dense detection, and visual perception for Multimodal Large Language Models (MLLMs). The training code and the pre-trained models have been released at <https://github.com/deepglint/MVT>.

1. Introduction

The proliferation of large-scale image-text data on the Internet has driven remarkable advances in visual representation learning. Pioneering models such as CLIP [50], DNF5B [21], OpenCLIP [27], EVA-CLIP [60], AIM [19], ViTamin [12], and SigLIP [80] have effectively leveraged image-text alignment at scale to learn highly robust and generalizable visual features. As a result, these visual encoders have become foundational components in a wide range of computer vision applications, paving the way for

more advanced MLLMs that integrate visual understanding with advanced language reasoning [32, 33, 38].

However, these models generally struggle to capture the underlying semantic structure of the training data (as shown in Fig. 1). This shortcoming arises from instance-wise contrastive learning, which treats samples from different instances as negative pairs, regardless of their semantic similarity. Furthermore, in vision-language contrastive learning, pairs that involve Optical Character Recognition (OCR) often cause the visual encoder to focus primarily on spotting text. As a result, high-level visual semantics in the images are overshadowed, leading to reduced performance on downstream zero-shot tasks that require object-centric and scene-centric understanding [49].

To overcome the limitations of instance discrimination, cluster discrimination methods such as DeepCluster [10], SeLa [5], SwAV [11], UNICOM [3], and MLCD [4] have been proposed. These approaches jointly learn cluster assignments and image embeddings, encouraging similar instances to be grouped together, and thereby better capturing the underlying semantic structure of the data. However, most cluster discrimination methods rely on assigning one or multiple pseudo-labels to each image, which limits their ability to learn localized region-level representations.

Recent works have sought to achieve vision-language alignment at the local region level to support dense prediction tasks. RegionCLIP [83] leverages the CLIP model to match image regions with template captions, pretraining the model to align region-text pairs in the feature space. CLIM [70] creates mosaics by combining multiple images, treating each image as a pseudo region, and then extracts features for each region. A contrastive loss is applied to ensure that features of each pseudo region are similar to their corresponding text embeddings and dissimilar to others. Nevertheless, these approaches rely on the availability of descriptive text for regions, which introduces additional constraints and limits scalability to larger datasets.

In this work, we construct region-specific datasets tai-

*corresponding author, † equal contribution

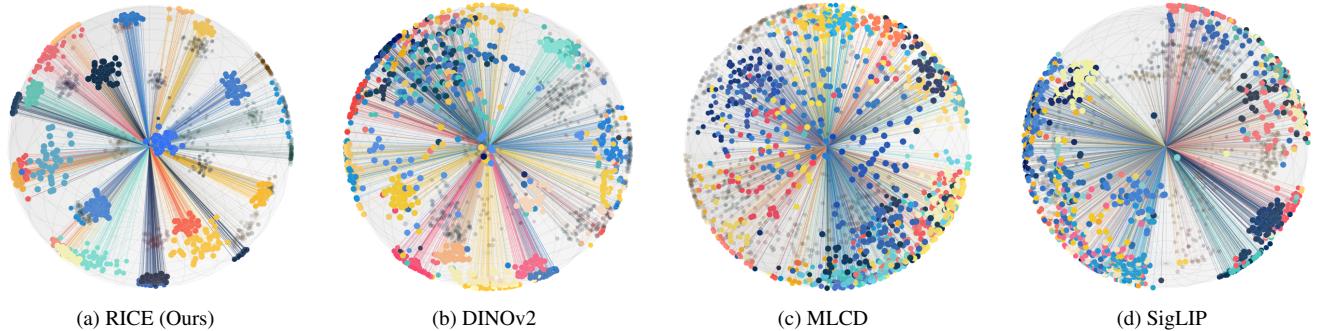


Figure 1. Visualization of object feature distributions from the COCO test dataset via t-SNE projection onto a spherical manifold.

lored for both region-aware object and OCR tasks. We propose a unified learning framework based on cluster discrimination that jointly optimizes region-level object recognition and OCR capabilities. Within this framework, object regions are formulated as single-label classification tasks, while OCR regions are addressed as multi-label classification problems. Specifically, we first curate a large-scale candidate region dataset. For object regions, we utilize SAM [30] to generate fine-grained mask regions for each image and extract CLIP features from these regions. Semantic labels are then assigned to each region using the k -means algorithm. For OCR regions, we employ PaddleOCR [18] to extract text data and generate corresponding OCR labels through tokenization.

To fully leverage the constructed region-specific datasets and enhance region-aware representation learning, we propose the RICE model, which integrates standard Transformer layers with novel Region Transformer layers. This architecture strengthens region-level representations while maintaining robust global feature learning. Additionally, we introduce an efficient region-cluster classification loss that unifies object and OCR learning within a single framework, enabling scalable distributed training on large-scale datasets. Extensive experiments show that RICE consistently outperforms previous methods across a range of downstream tasks, including segmentation, dense detection, and visual perception for MLLMs. In summary, our main **contributions** are as follows:

- We **propose** the RICE model for region-aware representation learning, introducing the Region Transformer layer as a core module for extracting rich regional semantic information.
- We **design** an efficient region-based cluster discrimination loss that unifies object and OCR learning within a single classification framework, enabling scalable training on large-scale datasets.
- We **conduct** extensive experiments, demonstrating the effectiveness of our approach across multiple downstream tasks, including segmentation, dense detection, and visual perception for MLLMs.

2. Related Work

Image Representation Learning. Existing image representation learning methods can be categorized into two approaches: (1) instance-based discrimination and (2) cluster-based discrimination. Instance-based methods, exemplified by CLIP [23, 50, 74, 75] and SigLIP [80], leverage a simple yet effective strategy that utilizes diverse label forms across different granularities for individual images. DINOv2 [48] introduces a self-supervised framework that does not require labels, achieving state-of-the-art results across a wide range of vision tasks. UNIT [85] integrates lightweight language and vision decoders, enabling robust text prediction while mitigating catastrophic forgetting in visual encoding. However, a key limitation of instance-wise contrastive learning is that it treats all different instances as negative pairs, which restricts its capacity to capture richer semantic relationships in the data. To address this limitation, cluster discrimination approaches such as UNICOM [3] and MLCD [4] have been proposed. These methods group multiple instances into clusters, drawing visually similar samples closer together and thus better capturing the underlying semantic structure. Nonetheless, most cluster discrimination methods assign one or more pseudo-labels to each image, which can hinder their ability to learn effective region-level representations.

Region Representation Learning. Early object detection methods [53, 54] typically rely on manually labeled regions to train their models. To alleviate the substantial cost of manual annotation, PreDet [51] extends self-supervised learning to region representation by enforcing similar embeddings for highly overlapping bounding boxes across different views of the same image, while contrasting them with non-overlapping regions. RegionCLIP [83] leverages the CLIP model to align image regions with template captions, pretraining the model to map region-text pairs into a shared feature space. GLIP [36] exploits large-scale image-text pairs by generating grounding boxes through self-training, thereby producing semantically rich representations. CLIM [70] constructs mosaics from multiple images, treating each image as a pseudo-region to facilitate

training. Distinct from these approaches, we propose leveraging region cluster centers as supervision signals. This strategy enhances semantic richness and facilitates scalable training by removing the reliance on descriptive textual annotations.

Multimodal Large Language Model. Motivated by the reasoning capabilities of Large Language Models (LLMs) [6, 15, 62, 63], recent research has explored transferring these abilities to the vision domain through the development of multimodal LLMs [7, 65]. Approaches such as Flamingo [1] and BLIP [34, 35] connect pre-trained vision encoders with frozen language models to enable multimodal understanding, while LLaVA [32, 38] achieves efficient vision-language alignment via instruction tuning. Recent advances in region-aware multimodal LLMs, such as MiniGPT-4 [84] and LLaVA-G [81], incorporate region information in textual form, relying primarily on the language decoder for positional interpretation. LISA [31] pioneered reasoning-based segmentation methods, and GLaMM [52] introduced new datasets and techniques for region-level captioning and segmentation. Concurrently, works such as [55, 68, 69, 73, 79] investigate integrating referring segmentation with conversational interfaces through instruction tuning. However, performance on complex region understanding tasks remains limited by the region representation capabilities of the image encoder.

3. Methodology

In this paper, we enhance the semantic object perception and OCR capabilities of visual representation models. To this end, we first construct region-aware datasets (Sec. 3.1), providing semantically rich supervision. We then introduce a Region Transformer layer (Sec. 3.2) to effectively leverage these data. Finally, we employ a region cluster discrimination loss (Sec. 3.3) that unifies general object recognition and OCR within a single classification framework.

3.1. Region Data Curation

Region-Aware Object Data Curation. To construct region-aware object data, we first sample images from the LAION2B [57], COYO700M [8], and SAM1B [30] datasets, ensuring each image has a minimum edge length of 336 pixels. For LAION2B and COYO700M, we apply the SAM model to generate fine-grained mask regions, while for SAM1B, we retain the original annotated regions. We further filter candidate bounding boxes to those with a minimum edge length of 128 pixels, resulting in a dataset comprising 400 million images and 2 billion candidate regions. Following UNICOM [3], we extract features from these regions using the CLIP model and apply the k -means algorithm to cluster them into one million semantic centers, $\mathcal{C} = \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$. The semantic label for each region is then assigned by nearest-neighbor matching to the cluster

center:

$$\mathbf{y}_{i,j}^{object} = \arg \min_{k \in [1, K]} \|\mathbf{f}_{i,j} - \mathbf{c}_k\|_2, \quad (1)$$

where $\mathbf{f}_{i,j} = \mathcal{F}_{CLIP}(\mathbf{r}_{i,j})$ denotes the CLIP feature embedding of region $\mathbf{r}_{i,j}$, which is the j -th semantic region in the i -th image. Overall, the clustering and label assignment step is relatively efficient. We implemented and parallelized this process using hierarchical k-means with the Faiss GPU library [17], completing the task in approximately 10 hours on 64 graphics cards.

Region-Aware OCR Data Curation. For OCR data, we use PaddleOCR [18] to extract text information from the LAION2B and COYO700M datasets, retaining only entries with a confidence score above 0.7. This yields a dataset of 50 million images and 400 million candidate regions. We then tokenize the extracted text for each region to obtain the corresponding OCR labels, $\mathbf{y}_{i,j}^{ocr} = \text{tokenizer}(\mathbf{t}_{i,j})$, where $\mathbf{t}_{i,j}$ denotes the text associated with region $\mathbf{r}_{i,j}$.

The final supervision signal combines semantic object cluster labels (\mathbf{y}^{object}) and OCR-derived labels (\mathbf{y}^{ocr}), depending on the selected region. The integration of these labels into the region-aware cluster discrimination framework is detailed in Sec. 3.3.

3.2. Region-aware Representation

Our framework is designed to achieve region-aware representation in a single inference pass. Unlike current visual encoders such as CLIP [50] and DINOv2 [48], our approach integrates standard Transformer layers for global semantic understanding with specialized Region Transformer layers for extracting regional semantics.

Region Sampling. To handle the varying number and size of regions in each image, we employ a balanced sampling strategy that standardizes the number of regions to N for efficient data loading. If the original number of regions $|\mathcal{R}_i|$ exceeds N , we randomly sample N regions from the index set $\mathcal{I}_{\mathcal{R}_i}$. Conversely, if $|\mathcal{R}_i| < N$, we retain all existing regions and augment the set by randomly resampling from $\mathcal{I}_{\mathcal{R}_i}$ until N regions are reached. This adaptive strategy ensures a consistent computational cost while preserving all available information when the region count is low.

Region Attention Layer. We adopt a sampling strategy that ensures a consistent number of regions per image. However, since regions vary in spatial size, the number of tokens per region remains inconsistent, complicating batch processing and hindering training scalability. To overcome this, we introduce a region attention layer as illustrated in Fig. 3. The proposed region attention layer applies a region-specific visibility mask to restrict attention to tokens within each designated region. This masking mechanism allows regions of different sizes to be efficiently grouped within the same batch, thereby facilitating scalable training.

Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we first apply a

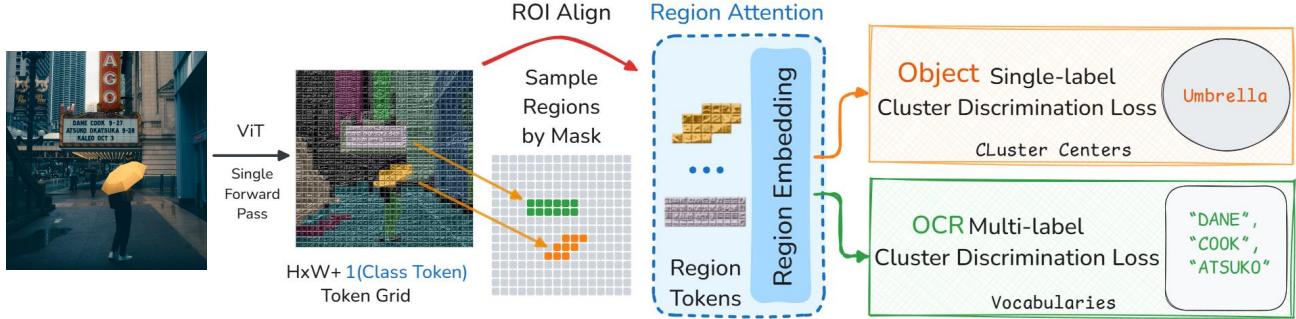


Figure 2. Overview of our unified semantic region understanding framework. Our approach efficiently processes diverse semantic regions within the image using a single forward pass. The model jointly captures both general visual semantics (objects) and OCR semantics (texts), seamlessly integrating them into a unified representation.

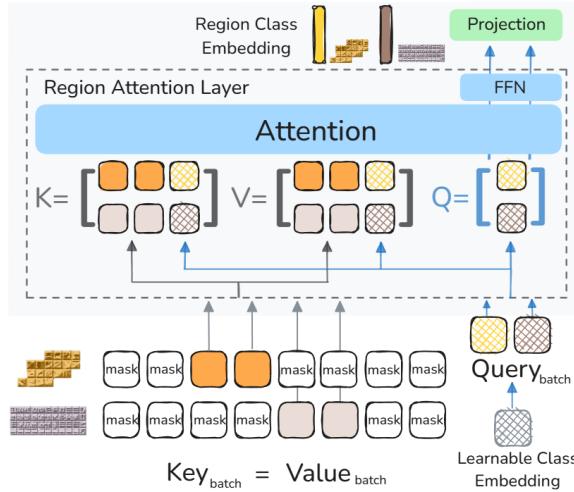


Figure 3. The region attention module processes batches of size-varying regions by leveraging a region-specific visibility mask and produces fixed-length region class embeddings. This approach enables efficient scaling of training while preserving representation fidelity.

2D convolutional layer to compute patch tokens $\mathbf{z} = \{\mathbf{z}_{cls}, \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\} \in \mathbb{R}^{(N+1) \times D}$, where $N = \frac{H}{p} \times \frac{W}{p}$ and p is the patch size. These tokens are then fed into Vision Transformer layers to produce global semantic embeddings $\mathcal{E} = \{e_{cls}, e_1, e_2, \dots, e_N\}$. Subsequently, we input these embeddings into the proposed Region Transformer layers to extract regional semantics. By employing standard global attention in the initial layers and region-aware attention in the later layers, our approach effectively balances the additional training overhead introduced by region-aware attention (whose batch size scales with N), while still preserving global contextual information.

As illustrated in Fig. 3, our proposed Region Attention Layer differs from the standard Vision Transformer by introducing a novel mask-guided region attention mechanism. To enable efficient parallel processing of all regions, we employ a region-specific visibility mask \mathcal{M} , which assigns a

value of 0 to tokens within the region and $-\infty$ to those outside the region. This visibility mask \mathcal{M} enforces region-constrained attention and facilitates batching across regions of varying sizes. The region attention operation is mathematically defined as:

$$\mathcal{R}_{batch} = \sigma \left(\frac{\mathbf{Q}_{batch} \mathbf{K}_{batch}^T}{\sqrt{d_k}} + \mathcal{M} \right) \mathbf{V}_{batch}, \quad (2)$$

where $\sigma(\cdot)$ denotes the softmax normalization function. This formulation enables the extraction of multiple region class embeddings in a single forward pass, optimizing computational efficiency while preserving representation fidelity. In terms of dimensions, the Query matrix \mathbf{Q}_{batch} has the shape (L, B, D) , where L is the number of regions processed, B is the batch size, and D is the embedding dimension. The Key and Value matrices, \mathbf{K}_{batch} and \mathbf{V}_{batch} , include all tokens of the original image. After the attention operation, the output \mathcal{R}_{batch} retains the same dimensions as the Query, yielding a (L, B, D) tensor that provides fixed-length representations for all regions.

3.3. Region Cluster Discrimination

After obtaining the embedding representation for each region, we introduce two specialized loss functions: (1) Object Region Loss for general visual recognition, and (2) OCR Region Loss for text recognition.

Object Region Loss. In the Object Region Loss, each region is associated with a single positive class center selected from a pool of one million class centers. During training, the embeddings are encouraged to move closer to their corresponding positive class center while simultaneously being pushed away from a randomly sampled subset of negative class centers [3]. The Object Region Loss is mathematically defined as follows:

$$\mathcal{L}_{object} = \log(1 + \exp(-sim(\hat{\mathbf{y}}_{i,j}, \mathbf{y}_{i,j}^{object}))) \quad (3)$$

$$+ \log(1 + \sum_{j \in \Omega_n^{object}} \exp(sim(\hat{\mathbf{y}}_{i,j}, \mathbf{y}_{i,j}^{object}))),$$

where $sim(\cdot)$ denotes the similarity function between two region features, and $\hat{\mathbf{y}}_{i,j}$ represents the feature embedding for the j -th region in the i -th image. Each region has a single positive class center, denoted as $\mathbf{y}_{i,j}^{object}$, while multiple negative class centers are randomly sampled from the set of negative centroids Ω_n^{object} .

OCR Region Loss. In the OCR Region Loss, positive classes are directly obtained from the token embeddings of the text within each OCR region. Each token embedding serves as a positive class center, resulting in multiple positive classes per OCR region. Consequently, the OCR Region Loss is designed to accommodate the intrinsic correspondence between text regions and multiple positive class centers (M) [4]. The OCR loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{ocr} = & \log(1 + \sum_{j \in \Omega_p^{ocr}} \exp(-sim(\hat{\mathbf{y}}_{i,j}, \mathbf{y}_{i,j}^{ocr}))) \quad (4) \\ & + \log(1 + \sum_{j \in \Omega_n^{ocr}} \exp(sim(\hat{\mathbf{y}}_{i,j}, \mathbf{y}_{i,j}^{ocr}))),\end{aligned}$$

where the negative labels are sampled from all other token embeddings, denoted as Ω_n^{ocr} . As previously noted, token embeddings computed within the same OCR region are selected as positives, denoted by Ω_p^{ocr} .

When conflicting classes are present, they can generate incorrect gradient signals that degrade model performance. To address this, we employ the random selection strategy [2] to efficiently construct a subset of negative class centers. Specifically, negative classes are uniformly sampled as $\Omega_n = \Omega_{\text{uniform}}^k(1, 2, \dots, K \setminus \Omega_p)$, where $k = \lfloor K \times \rho \rfloor$ and $\rho \in (0, 1]$ controls the sampling density. This strategy offers three main benefits: it improves computational efficiency by reducing overhead, lowers the chance of including semantically similar classes as negatives thus reducing conflicting gradients, and promotes more stable model convergence. Experiments show that an appropriate sampling ratio ρ significantly enhances performance.

4. Experiments

4.1. Implementation Details

Pretraining Setup. We train our RICE models on LAION-2B [56], SAM1B [30], and COYO-700M [8], processing a total of 13 billion samples during the initial pretraining stage. Training is conducted with a global batch size of 32K distributed across 64 graphics cards. We utilize the AdamW [43] optimizer with a learning rate of 0.001 and a decoupled weight decay coefficient of 0.2. In this paper, we employ ViT-L/14 and ViT-B/16 architectures. For ViT-L/14, we initially train at a resolution of 224×224 , followed by fine-tuning at 336×336 , 378×378 , and 560×560 . For ViT-B/16, training starts at 224×224 and proceeds to 512×512 for fine-tuning. At higher resolutions, the learning rate

is reduced by an order of magnitude, and 1 billion samples are used for fine-tuning. Throughout all experiments, the number of classes (K) is fixed at one million. Following the margin-based classification approach [16], we apply L2 normalization to both the feature vectors and class centers, and set a margin of $m = 0.3$ for positive classes with a scale parameter of 64. The ratio of sampled negative class centers (ρ) is fixed at 0.1, as in previous work [3, 4].

Multimodal Setup. For our multimodal large language model evaluations, we adopt the LLaVA-NeXT [33, 39] framework to ensure experimental consistency. All training procedures strictly follow the original LLaVA-NeXT implementation, using the same pretraining datasets and instruction-tuning data. We utilize Qwen2.5-7B [26] as the language model backbone to avoid hyperparameter biases that could favor the OpenAI CLIP model in the original setup. This controlled experimental design enables a fair comparison when evaluating the performance of our vision encoders within multimodal systems.

Referring Image Segmentation Setup. For our referring image segmentation evaluation, we integrate our vision encoders into LLaVA-NeXT using Qwen2.5-7B as the language model backbone. Following LISA [31], we adopt their two-stage training approach, vision-language alignment followed by MLLM-Decoder training with SAM integration. We also employ the same training data mixture of semantic and referring segmentation datasets. Additionally, we incorporate the LLaVA-NeXT 740K instruction dataset [33]. For segmentation tasks, we introduce a specialized [SEG] token, whose embedding is transformed via an MLP adapter to generate SAM prompts.

4.2. RICE as a Vision Encoder for MLLMs

Performance of Multimodal Understanding. Based on Tab. 1, we comprehensively evaluate the RICE model across multiple vision backbones within leading multimodal large language model frameworks. At a resolution of 336px, RICE achieves substantial performance gains over the widely used CLIP model and consistently outperforms more complex models with higher input resolutions, such as SigLIP and DFNS5B. Notably, RICE delivers significant improvements on OCR-related tasks: on OCRBench, RICE surpasses CLIP-336px by +50 points and SigLIP-384px by +34 points, and on DocVQA, RICE achieves improvements of +3.98%, +5.68%, and +4.30% over the respective baselines. Similar trends are observed for InfoVQA and TextVQA. At higher resolutions (560px), RICE maintains its lead, outperforming SigLIPv2-560px by +2.92% on InfoVQA and +1.18% on DocVQA. Remarkably, RICE-560px attains a DocVQA score of 87.38%, exceeding the performance of Qwen2.5-VL’s specialized backbone (85.83%). Furthermore, across both LLaVA-NeXT and LLaVA-OneVision frameworks, RICE outper-

Model Configuration			OCR & Document Understanding								General Vision Understanding							
Method	Vision Tower	LLM	InfoVQA[46]	DocVQA[45]	CharQA[44]	TextVQA[59]	OCRBench[41]	OCRBenchV2[41]	LiveXivVQA[58]	OCR Avg	AID[29]	MMBEN[40]	MME ^{Cog} [77]	MME ^{Per} [77]	POPE[37]	RealworldQA[72]	MMStar[13]	Other Avg
LLaVA-NeXT Framework																		
CLIP [50]	ViT-L-14-336px	Qwen2.5-7B	38.88	75.21	66.52	62.47	525	22.95	47.35	52.27	73.15	74.57	384	1512	88.83	63.66	48.98	69.83
MLCD [4]	ViT-L-14-336px	Qwen2.5-7B	43.48	76.46	67.84	61.69	531	23.98	48.43	53.57	76.98	76.37	433	1598	88.69	61.05	50.98	72.31
AIMv2 [22]	ViT-L-14-336px	Qwen2.5-7B	35.44	77.19	72.72	65.85	572	23.92	47.28	54.23	75.36	78.61	386	1500	88.41	62.22	50.16	70.58
RICE	ViT-L-14-336px	Qwen2.5-7B	45.23	79.19	72.34	65.89	575	24.14	48.94	56.18	77.88	76.55	437	1613	88.49	63.05	51.75	73.03
↑ RICE-336px vs AIMv2			+9.79	+2.00	-0.38	+0.04	+3	+0.22	+1.66	+1.95	+2.52	+2.06	+51	+113	+0.08	+0.83	+1.59	+2.45
DFN5B [21]	ViT-H-14-378px	Qwen2.5-7B	38.59	70.87	64.36	59.40	473	21.88	46.17	49.80	73.51	73.37	366	1537	88.62	59.87	49.14	68.91
UNIT [85]	ViT-H-14-378px	Qwen2.5-7B	42.83	77.21	69.94	65.41	559	22.94	46.43	54.38	74.24	73.02	354	1554	88.33	61.00	46.46	68.54
SigLIP [80]	ViT-SO400M-14-384px	Qwen2.5-7B	41.38	76.71	69.28	64.74	554	23.97	48.42	54.27	76.17	76.98	369	1597	88.83	63.66	47.32	70.62
SigLIPv2 [66]	ViT-SO400M-14-384px	Qwen2.5-7B	43.66	79.14	70.16	66.23	587	25.35	48.59	55.98	76.98	77.14	373	1607	89.26	63.36	52.77	71.70
RICE	ViT-L-14-378px	Qwen2.5-7B	48.05	82.59	75.08	66.24	588	25.84	49.54	58.02	76.54	77.60	433	1580	89.14	62.88	51.20	72.63
↑ RICE-378px vs SigLIPv2			+4.39	+3.45	+4.92	+0.01	+1	+0.49	+0.95	+2.04	+0.44	+0.46	+60	-27	-0.12	-0.48	-1.57	+0.93
SigLIPv2 [66]	ViT-SO400M-16-560px	Qwen2.5-7B	50.23	86.20	77.40	70.23	627	26.47	52.94	60.88	76.98	76.46	428	1597	89.26	68.24	53.06	73.60
Qwen2.5-ViT [64]	ViT-H-14-560px	Qwen2.5-7B	55.91	85.83	78.84	73.70	662	26.84	53.44	62.97	78.76	78.44	496	1616	88.64	64.18	54.98	75.50
RICE	ViT-L-14-560px	Qwen2.5-7B	53.15	87.38	78.08	68.96	607	26.14	53.03	61.06	76.88	78.61	450	1585	88.94	65.10	50.47	73.46
↑ RICE-560px vs SigLIPv2-560px			+2.92	+1.18	+0.68	-1.27	-20	-0.33	+0.09	+0.18	-0.10	+2.15	+22	-12	-0.32	-3.14	-2.59	-0.14
LLaVA-OneVision Framework																		
SigLIP [80]	ViT-SO400M-14-384px	Qwen2-7B	68.80	87.50	80.00	78.50	697	27.10	58.44	67.15	81.40	80.80	418	1580	87.20	66.30	61.70	75.15
RICE	ViT-L-14-378px	Qwen2-7B	73.57	91.80	82.71	79.59	749	32.05	60.96	70.80	84.57	81.75	564	1601	88.83	68.94	63.88	80.29
↑ RICE vs SigLIP (OneVision)			+4.77	+4.30	+2.71	+1.09	+52	+4.95	+2.52	+3.65	+3.17	+0.95	+146	+21	+1.63	+2.64	+2.18	+5.14

Table 1. Comprehensive performance comparison of RICE with state-of-the-art vision encoders. For all experiments within the LLaVA-NeXT framework, we adopt a high-resolution tiling strategy: each input image is divided into a $2 \times 2 + 1$ grid of crops, where each crop matches the pre-training resolution of the backbone model (e.g., 336px, 378px, or 560px). The scores for OCRCbench, MME^{Cog}, and MME^{Per} are normalized (divided by 10, 8, and 20, respectively) to scale all results to a 0–100 range.

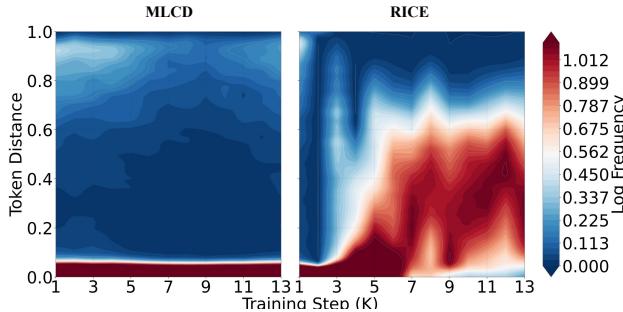


Figure 4. Token distance distributions observed during the training of MLCD and RICE models.

forms CLIP and SigLIP on most benchmarks, with particularly strong gains in perception tasks. Collectively, these results establish RICE as a robust and versatile vision encoder, achieving substantial improvements on OCR-centric benchmarks, maintaining strong general vision performance, and delivering results competitive with specialized models trained on extensive datasets.

Performance of Referring Segmentation. Table 2 presents an in-depth comparison of various vision-language model configurations using both the LLaVA-1.5 and LLaVA-NeXT (1.6) architectures. Within the LLaVA-1.5 framework, the RICE vision tower combined with Vicuna-7B consistently outperforms the standard CLIP vision en-

coder, achieving improvements of 1.4%, 1.2%, and 2.8% on the refCOCO val, testA, and testB splits, respectively. Notably, in the more advanced LLaVA-NeXT framework with Qwen2.5-7B, RICE significantly surpasses the baseline MLCD method across all benchmarks. Specifically, RICE achieves relative gains of 0.7% on refCOCO val, 0.7% on refCOCO testA, and 1.5% on refCOCO testB. This significant performance improvement is attributed to RICE’s ability to more accurately express local semantic information. To further validate this point, we calculated pairwise distances between different image tokens based on the COCO dataset. As illustrated in Fig. 4, RICE demonstrates superior capability to differentiate between visual tokens during training compared to MLCD, thereby enabling more precise object perception within the image.

4.3. Detection Probe

We evaluate RICE against several state-of-the-art pre-trained vision encoders across multiple benchmarks. All evaluations are conducted using the Cascade Mask R-CNN framework [9, 24] implemented in Detectron2 [71]. As summarized in Table 3, experiments are performed on the COCO (80 classes) and LVIS (1203 classes) validation sets. Our feature pyramid is constructed directly from frozen backbone embeddings, employing max-pooling and unpooling operations to generate multi-scale feature maps

VisionTower	LLM	Method	refCOCO [28]			refCOCO+ [28]			refCOCOg [78]	
			val	testA	testB	val	testA	testB	val	test
Previous Methods										
CLIP	Vicuna-7B [15]	GLaMM [52]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9
CLIP	Vicuna-7B [15]	VisionLLMv2 [69]	79.2	82.3	77.0	68.9	75.8	61.8	73.3	74.8
CLIP	Vicuna-7B [15]	LLaVA-G-7B [81]	77.1	-	-	68.8	-	-	71.5	-
CLIP	LLaMA2-13B [62]	GSVA [73]	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0
CLIP	LLaMA2-13B [62]	PixelLM-7B [55]	73.0	-	-	66.3	-	-	69.3	-
ConvNext-L [27, 42]	InternLM2-7B [61]	OMG-LLaVA [82]	77.2	79.8	74.1	68.7	73.0	61.6	71.7	71.9
InternViT2.5-300M [14]	InternLM2.5-7B [14]	Sa2VA [79]	81.6	-	-	76.2	-	-	78.7	-
InternViT2.5-6B [14]	InternLM2.5-20B [14]	Sa2VA [79]	82.5	-	-	78.8	-	-	79.7	-
LLaVA-1.5 Framework										
CLIP	Vicuna-7B [15]	LISA [31]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
RICE	Vicuna-7B [15]	LISA [31]	76.3	80.3	75.1	67.4	72.7	60.6	69.0	73.4
Avg: +2.00 ↑ Improvement over CLIP			+1.4	+1.2	+2.8	+2.3	+1.9	+2.5	+1.1	+2.8
LLaVA-NeXT Framework										
CLIP	Qwen2.5-7B	LISA [31]	81.8	84.0	79.1	76.6	80.5	70.9	77.3	78.5
MLCD	Qwen2.5-7B	LISA [31]	82.8	84.6	80.2	77.4	81.6	73.1	78.5	79.7
RICE	Qwen2.5-7B	LISA [31]	83.5	85.3	81.7	79.4	82.8	75.4	79.8	80.4
Avg: +2.45 ↑ Improvement over CLIP			+1.7	+1.3	+2.6	+2.8	+2.3	+4.5	+2.5	+1.9
Avg: +1.30 ↑ Improvement over MLCD			+0.7	+0.7	+1.5	+2.0	+1.2	+2.3	+1.3	+0.7

Table 2. Performance comparison of referring image segmentation across vision-language models. Results are reported as IoU scores (%) on the refCOCO, refCOCO+, and refCOCOg benchmarks. Our RICE vision encoder consistently outperforms all competing approaches, achieving state-of-the-art results across all benchmarks when integrated with Qwen2.5-7B in the LLaVA-NeXT framework.

Configuration			COCO		LVIS		Roboflow100 Benchmarks							
Method	Arch	Res	Det AP	Seg AP	Det AP	Seg AP	Aerial	Video Games	Microscopic	Under Water	Documents	Electro-magnetic	Real world	AVG
DINOv2	ViT-B/14	518	31.6	24.3	18.7	14.1	2.3	14.3	10.6	19.9	18.8	15.3	26.8	15.4
SigLIP	ViT-B/16	512	35.0	28.1	21.8	17.3	9.4	29.5	20.0	29.4	23.4	18.6	38.0	24.1
MLCD	ViT-B/16	512	35.6	28.6	22.1	17.8	11.4	19.9	14.9	21.0	13.3	15.8	25.0	17.3
RICE	ViT-B/16	512	38.9	31.5	26.5	21.4	14.9	31.7	23.4	30.7	27.0	18.7	39.1	26.5

Table 3. Performance comparison of vision encoders across multiple benchmarks, grouped by model size. Results are reported as Average Precision (%) on COCO Detection, COCO Segmentation, LVIS Detection, LVIS Segmentation, and the Roboflow100-VL benchmark across various domains. Our proposed RICE vision encoder exhibits superior performance, consistently outperforming existing methods across these diverse evaluation settings.

at resolutions of $0.25\times$, $0.5\times$, $1\times$, $2\times$, and $4\times$. RICE achieves superior performance across all evaluation metrics. On COCO, RICE attains 38.9% detection AP and 31.5% segmentation AP, surpassing the strongest baseline, SigLIP, by +3.9% and +3.4%, respectively. On LVIS, RICE achieves 26.5% detection AP and 21.4% segmentation AP, representing improvements of +4.7% and +4.1% over SigLIP. On the Roboflow100 benchmarks, RICE demonstrates exceptional generalization across diverse domains, achieving a 26.5% average performance and notable gains in aerial imagery (+5.5%) and microscopic analysis (+3.4%). These consistent improvements on both natural images (COCO/LVIS) and specialized domains (Roboflow100) highlight RICE’s superior representational quality and its effectiveness for a wide range of computer vision applications.

4.4. Tracking Probe

We evaluate the temporal matching capability of local features within the general video object tracking framework

OTrack [76], adopting an attention probing approach to compare the four pre-trained models. Specifically, we insert two standard vision transformer blocks between the frozen backbone and the prediction head to enhance information exchange between the template and search images, while only passing the search image features to the position prediction network. For GOT-10k [25], models are trained for 100 epochs on the GOT-10k training set. For evaluations on LaSOT [20], TrackingNet [47], and TNL2K [67], we train for 300 epochs using the combined training sets from COCO, TrackingNet, GOT-10k, and LaSOT. As shown in Tab. 4, RICE achieves the best performance across all metrics on the above four datasets.

To further interpret the model’s behavior, we visualize semantic features by projecting token representations from the ViT-B/16 model onto RGB channels via PCA. The coherence of color patterns across frames demonstrates the model’s ability to dynamically maintain focus on salient objects while adapting to movement and contextual changes. In Fig. 5, we visualize frames from diverse videos, where

Method	LaSOT			TrackingNet			GOT-10k			TNL2k		
	Suc.	Pre.	Norm. Pre.	Suc.	Pre.	Norm. Pre.	AO	SR-0.5	SR-0.75	Suc.	Pre.	Norm. Pre.
DINOv2	55.11	54.99	65.52	71.20	64.70	77.70	53.60	64.90	35.50	41.95	36.03	57.40
SigLIP	55.52	56.16	65.33	72.60	66.70	78.70	53.50	63.10	35.40	43.90	39.06	59.03
MLCD	58.05	60.75	68.31	75.30	70.20	80.80	53.80	62.80	39.70	45.22	40.62	60.64
RICE	60.24	63.16	69.66	76.30	71.80	81.30	55.40	63.50	41.60	45.70	41.95	61.18

Table 4. Performance comparison of different vision encoders on video tracking benchmarks. Results are reported on LaSOT, TrackingNet, GOT-10k, and TNL2k datasets. Our proposed RICE vision encoder demonstrates superior performance across all metrics.

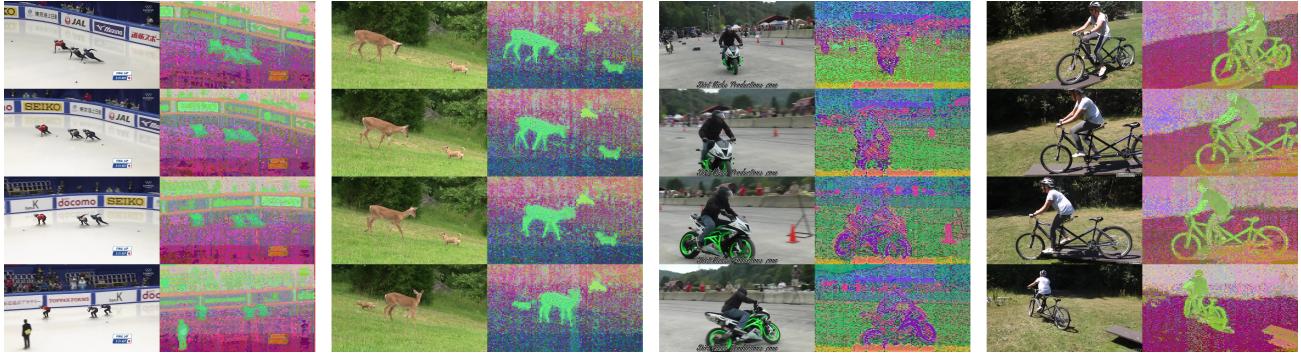


Figure 5. Tracking PCA. Using 2048-resolution images as input to a ViT-B/16 model, we project token features onto RGB channels via PCA to visualize the semantic structure. Sequential frames (arranged vertically) illustrate the evolution of model attention, consistently highlighting salient objects across time. The visualization reveals stable color patterns for tracked entities such as ice skaters, deers, motorcyclists, and cyclists, demonstrating the model’s ability to maintain semantic focus throughout the sequence.

Testset	1	5	10	20	Testset	200K	500K	1M	2M	Testset	0.05	0.1	0.2	0.5	Testset	3	5	10	20
COCO-Det	31.4	34.5	35.2	34.7	COCO-Det	32.8	33.6	35.2	34.9	COCO-Det	38.8	35.2	35.1	33.7	DocVQA	53.1	54.3	54.4	53.0
COCO-Seg	20.7	27.1	28.6	27.2	COCO-Seg	24.3	26.8	28.6	27.9	COCO-Seg	27.3	28.6	28.4	27.4	ChartQA	62.7	62.2	63.0	62.9
LVIS-Det	17.5	21.2	23.1	22.4	LVIS-Det	18.9	21.6	23.1	24.0	LVIS-Det	22.1	23.1	22.4	21.6	InfoVQA	37.3	37.5	38.2	38.4
LVIS-Seg	14.3	17.3	19.9	19.0	LVIS-Seg	15.4	17.8	19.9	18.7	LVIS-Seg	19.8	19.9	19.1	18.6	OCRbench	391	392	403	401

(a) Sampled boxes (N).

(b) Cluster centers (K).

(c) Negative sampling rate (ρ).

(d) Positive OCR labels (M).

Table 5. Ablation studies on key hyperparameters using the ViT-B/16 backbone and 10% of the pretraining dataset to investigate optimal settings for detection, segmentation, and OCR tasks.

each vertical sequence depicts consecutive frames, the original images on the left and their corresponding PCA visualizations on the right. RICE keeps stable attention on tracked entities such as ice skaters, deers, motorcyclists, and cyclists, not only confirming the robustness of the model’s object tracking capabilities but also illustrating the value of PCA-based visualization for revealing semantic feature distributions and temporal attention dynamics.

4.5. Ablation Study

We conducted extensive ablation studies on key hyperparameters using the ViT-B/16 backbone and 10% of the pretraining dataset, as presented in Tab. 5a–5d. Specifically, Tab. 5a demonstrates that utilizing 10 object bounding boxes yields the best performance across all detection and segmentation benchmarks. For cluster centers (Tab. 5b), we observe optimal performance with 1–2 million centers, where COCO benchmarks peak at 1M and LVIS-Det achieves its highest results at 2M. As shown in Tab. 5c, lower negative sampling rates lead to superior outcomes,

with COCO-Det attaining the best performance at a rate of 0.05, while other metrics perform best at 0.1. Our experimental framework also includes analyses on the impact of positive OCR labels, as reported in Tab. 5d.

5. Conclusions

In this paper, we propose a novel Region-Aware Cluster Discrimination approach to advance region-aware object recognition and OCR capabilities. We first construct a billion-scale candidate region dataset and introduce the RICE model, which incorporates a Region Transformer layer to capture regional semantic information. Furthermore, we develop an efficient region-based classification loss that jointly supports object and OCR learning, enabling scalable and distributed training on large-scale data. Extensive experiments demonstrate that RICE consistently outperforms prior methods across a range of downstream tasks, including segmentation, dense detection, and visual perception for MLLMs.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022. 3
- [2] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing Two Birds with One Stone: Efficient and Robust Training of Face Recognition CNNs by Partial FC. In *CVPR*, 2022. 5
- [3] Xiang An, Jiankang Deng, Kaicheng Yang, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicorn: Universal and Compact Representation Learning for Image Retrieval. In *ICLR*, 2023. 1, 2, 3, 4, 5
- [4] Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, and Jiankang Deng. Multi-label Cluster Discrimination for Visual Representation Learning. In *ECCV*, 2024. 1, 2, 5, 6
- [5] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-Labelling via Simultaneous Clustering and Representation Learning. In *ICLR*, 2020. 1
- [6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. 2023. 3
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv:2308.12966*, 2023. 3
- [8] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-Text Pair Dataset, 2022. 3, 5
- [9] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*, 2018. 6
- [10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, 2018. 1
- [11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, 2020. 1
- [12] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. ViTamin: Designing Scalable Vision Models in the Vision-language Era. In *CVPR*, 2024. 1
- [13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *NeurIPS*, 2024. 6
- [14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv:2412.05271*, 2024. 7
- [15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing.
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. 2023. 3, 7
- [16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5
- [17] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazare, Maria Lomeli, Lucas Hosseini, and Herve Jegou. The Faiss library. *arXiv:2401.08281*, 2025. 3
- [18] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. PP-OCR: A Practical Ultra Lightweight OCR System. *arXiv:2009.09941*, 2020. 2, 3
- [19] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable Pre-training of Large Autoregressive Image Models. *PMLR*, 2024. 1
- [20] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 7
- [21] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data Filtering Networks. In *ICLR*, 2024. 1, 6
- [22] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Heldmann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *CVPR*, 2025. 6
- [23] TIANCHENG GU, KAICHENG YANG, XIANG AN, ZIYONG FENG, DONGNAN LIU, WEIDONG CAI, AND JIANKANG DENG. RWKV-CLIP: A robust vision-language representation learner. In *EMNLP*, 2024. 2
- [24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 6
- [25] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *TPAMI*, 2019. 7
- [26] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2.5-Coder Technical Report. *arXiv:2409.12186*, 2024. 5
- [27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. 2021. 1, 7
- [28] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 7
- [29] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. In *ECCV*, 2016. 6

- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, 2023. 2, 3, 5
- [31] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 3, 5, 7
- [32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv:2408.03326*, 2024. 1, 3
- [33] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. In *ICLR*, 2025. 1, 5
- [34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022. 3
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 3
- [36] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-training. In *CVPR*, 2022. 2
- [37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 2023. 6
- [38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024. 1, 3
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 5
- [40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281*, 2023. 6
- [41] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRCBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 2024. 6
- [42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 7
- [43] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2018. 5
- [44] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL Findings*, 2022. 6
- [45] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A Dataset for VQA on Document Images. In *WACV*, 2021. 6
- [46] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographiccvqa. In *WACV*, 2022. 6
- [47] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 7
- [48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINoV2: Learning Robust Visual Features without Supervision. *PMLR*, 2024. 2, 3
- [49] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Miyaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. In *CVPR*, 2023. 1
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1, 2, 3, 6
- [51] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Pre-Det: Large-Scale Weakly Supervised Pre-Training for Detection. In *ICCV*, 2021. 2
- [52] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdellrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *CVPR*, 2024. 3, 7
- [53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. 2
- [54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 2
- [55] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel Reasoning with Large Multimodal Model. In *CVPR*, 2024. 3, 7
- [56] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv:2111.02114*, 2021. 5
- [57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An Open Large-Scale

- Dataset for Training Next Generation Image-Text Models. *arXiv:2210.08402*, 2022. 3
- [58] Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, et al. LiveXiv—A Multi-Modal Live Benchmark Based on Arxiv Papers Content. *arXiv:2410.10783*, 2024. 6
- [59] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *CVPR*, 2019. 6
- [60] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv:2303.15389*, 2023. 1
- [61] Intern Team. InternLM2 Technical Report. *arXiv:2403.17297*, 2024. 7
- [62] LLaMA Team. LLaMA2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*, 2023. 3, 7
- [63] Qwen Team. Qwen2 Technical Report. *ArXiv:2407.10671*, 2024. 3
- [64] Qwen Team. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025. 6
- [65] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *NeurIPS*, 2024. 3
- [66] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. 2025. 6
- [67] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, 2021. 7
- [68] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. LaSagnA: Language-based Segmentation Assistant for Complex Queries. *arXiv:2404.08506*, 2024. 3
- [69] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhui Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. VisionLLM v2: An End-to-End Generalist Multimodal Large Language Model for Hundreds of Vision-Language Tasks. In *NeurIPS*, 2024. 3, 7
- [70] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. CLIM: Contrastive Language-Image Mosaic for Region Representation. *arXiv:2312.11376*, 2023. 1, 2
- [71] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019. 6
- [72] XAI.ORG. Grok-1.5 Vision Preview, 2024. 6
- [73] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *CVPR*, 2023. 3, 7
- [74] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, 2023. 2
- [75] Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. Clip-cid: Efficient clip distillation via cluster-instance discrimination. In *AAAI*, 2025. 2
- [76] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 7
- [77] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *arXiv:2306.13549*, 2023. 6
- [78] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016. 7
- [79] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv:2501.04001*, 2025. 3, 7
- [80] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, 2023. 1, 2, 6
- [81] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models. In *ECCV*, 2024. 3, 7
- [82] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. OMG-LLaVA: Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding. In *NeurIPS*, 2024. 7
- [83] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*, 2022. 1, 2
- [84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 3
- [85] Yi Zhu, Zhou Yanpeng, Chunwei Wang, Yang Cao, Jianhua Han, Lu Hou, and Hang Xu. UNIT: Unifying Image and Text Recognition in One Vision Encoder. In *NeurIPS*, 2024. 2, 6