MultiModal Code-Switching: Interleaving Visual Objects into Language for Explicit Object-Level Alignment

Anonymous ACL submission

Abstract

Existing Multimodal Large Language Models (MLLMs) predominantly rely on image-text pairs for modality alignment pretraining, mapping global image representations to long textual descriptions. However, this image-level alignment suffers from *referential ambiguity*: models struggle to infer the correspondences between multiple visual objects and textual entities from the global representation, leading to data inefficiency and suboptimal semantic grounding. To address this, we propose MultiModal Code-Switching (MMCS), a novel pretraining paradigm that provides explicit object-level supervision. Inspired by the linguistic phenomenon of code-switching, MMCS interleaves vision and language by replacing textual entities with their corresponding visual objects, enforcing local vision-language grounding. We further develop a scalable data synthesis pipeline to generate a pretraining dataset of 773K samples with accurate object–entity correspondences. Experiments show that MMCS is highly data-efficient: with only 50K samples, it matches or surpasses models trained on 600K image–text pairs. Furthermore, MMCS consistently improves visual grounding and perception capabilities across varying model scales[1].

1 Introduction

Multimodal Large Language Models (MLLMs) have established a new state-of-the-art in vision-language understanding, demonstrating exceptional performance across visual question answering (Goyal et al., 2017; Hudson and Manning, 2019), document understanding (Mathew et al., 2021), and visual grounding (Kazemzadeh et al., 2014; Mao et al., 2016). The dominant architecture for these models (Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023; Bai et al., 2025b; Zhu et al., 2025) generally comprises a vision encoder, a Large Language Model (LLM) backbone, and a projector (e.g., MLP or Q-Former) that bridges the modality gap. To unify these components, the standard training paradigm follows a two-stage process: 1) modality alignment pretraining, mapping visual features into the LLM's semantic space via image-text pairs; and 2) multimodal supervised instruction tuning (SFT), optimizing the model for downstream task execution. Within this framework, modality alignment is foundational, as the fidelity of this alignment dictates the upper bound of the model's multimodal capabilities (Liu et al., 2023; McKinzie et al., 2024).

The prevailing consensus in recent research is to utilize dense image captions for modality alignment, providing detail-rich linguistic signals (Chen et al., 2024a,b; Li et al., 2024, 2025; Deitke et al., 2025; Xing et al., 2025). In this paradigm, the model encodes the image into a global image representation, and is then trained to predict a lengthy text sequence based on this global representation, thereby achieving alignment at the image level. However, natural images are inherently complex, often containing multiple objects and background elements. As highlighted in Figure 1 (right), standard dense caption datasets contain an average of 6.8 to 11.0 distinct entities per sample (Chen et al., 2024b; Onoe et al., 2024; Li et al., 2024; Garg et al., 2024). While the captions meticulously enumerate specific objects and attributes, the vision encoder and projector compress the entire scene into a generic, global representation.

This discrepancy leads to the issue of *referential ambiguity*: the model must implicitly infer the correspondence between specific visual regions and the corresponding textual phrases from global representations. From a computational perspective, this "many-to-many" mapping forces the model to rely on statistical co-occurrences rather than genuine semantic grounding. Consequently, this implicit alignment paradigm is highly data-inefficient,

[1]We provide an anonymous code repository at https://anonymous.4open.science/r/MMCS_anonymous-4C7A/

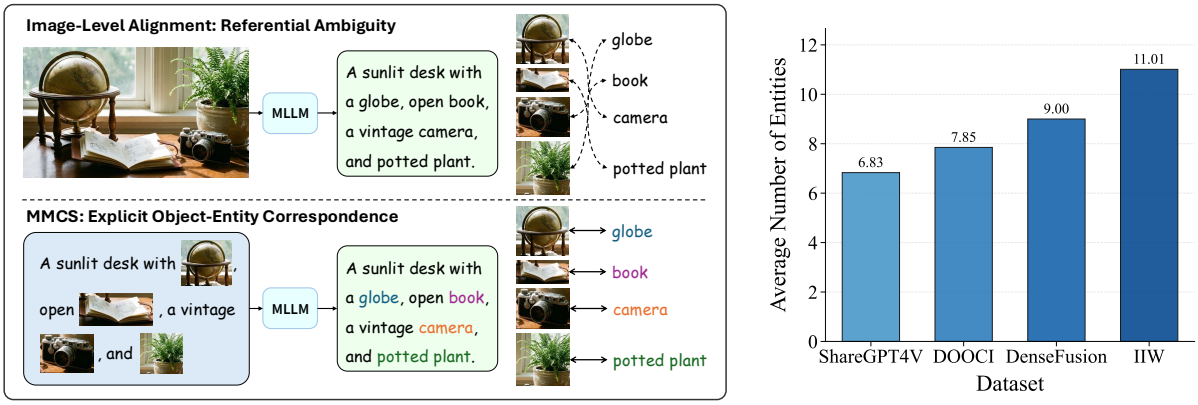


Figure 1: **Left**: Illustration of the referential ambiguity in standard image-level alignment (top), contrasted with our MultiModal Code-Switching (MMCS) paradigm (bottom). MMCS resolves ambiguity by replacing textual entities with their corresponding visual objects to provide explicit correspondence signals. **Right:** The large average number of entities in dense caption datasets highlights the complexity of natural images.

necessitating massive-scale datasets to learn robust object-entity associations (McKinzie et al., 2024; Dong et al., 2025). Our analysis in Section 5 substantiates this by revealing that models pretrained with standard image-caption pairs exhibit diffuse attention patterns and suboptimal representation consistency.

To address these limitations, we propose MultiModal Code-Switching (MMCS), a novel pretraining paradigm that introduces *explicit object-level supervision*. Inspired by the linguistic phenomenon of code-switching (Poplack, 1981; Thara and Poornachandran, 2018), MMCS treats vision and language as distinct "codes". Instead of relying solely on global image context, we create interleaved representations by substituting the embeddings of textual entities with the embeddings of their corresponding visual objects (Figure 1, left). By conditioning the generation of the immediate textual context directly on these local visual features, MMCS imposes a structural constraint that enforces explicit grounding of textual entities to corresponding visual regions. This eliminates the need for the model to infer correspondences from global representations, thereby facilitating object-level alignment efficiently.

To implement this, we develop a data synthesis pipeline to generate 773K high-quality samples with accurate object-entity correspondences. Given an image, our pipeline generates a detailed caption, extracts textual entities, and employs a grounding model to localize the corresponding visual objects. Empirically, MMCS demonstrates extraordinary data efficiency compared to standard image-level pretraining. With only 50K MMCS samples, our model achieves performance exceeding models pretrained on 600K standard image-caption pairs. Moreover, MMCS maintains consistent gains when scaling up both the dataset size and model capacity, yielding average improvements of 7.9% on visual grounding and 2.1% on perception-centric benchmarks. Further in-depth analysis confirms that these gains stem from higher-fidelity representation alignment and sharper attention distributions.

Our contributions are summarized as follows:

- We introduce MultiModal Code-Switching (MMCS), a novel pretraining paradigm that interleaves visual objects into text, shifting alignment from implicit image-level associations to **explicit object-level correspondences**.
- We develop a **scalable data synthesis pipeline** that generates 773K samples with precise object-entity correspondences, bypassing the need for manual annotation.
- We conduct extensive experiments across various model scales and vision encoders, demonstrating the effectiveness of MMCS. We further provide **mechanistic insights** into how MMCS enhances the internal feature space topology of MLLMs.

2 Related Works

Multimodal Large Language Models Current mainstream MLLMs adopt the ViT-MLP-LLM paradigm (An et al., 2025; Yue et al., 2025; Zhu

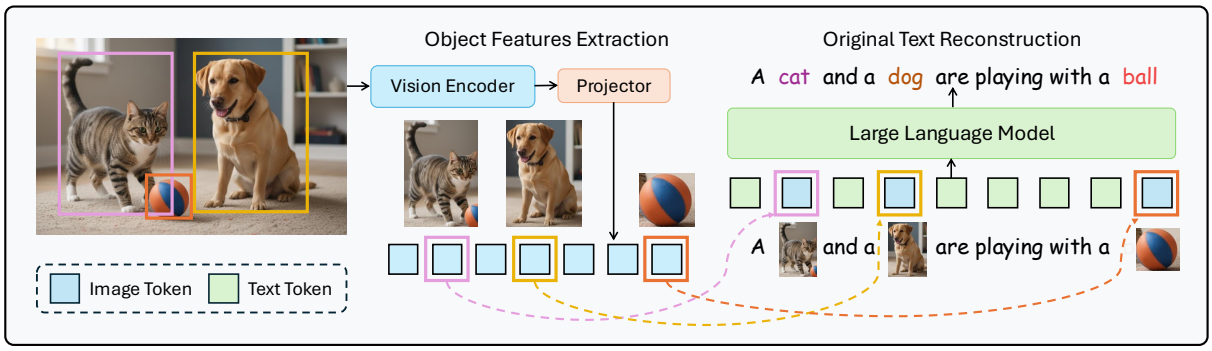


Figure 2: **Overview of MMCS pretraining paradigm.** We construct an interleaved image-text sequence by substituting textual entities with corresponding extracted visual objects. The generation of these entities and the subsequent context is conditioned on its visual counterparts, thereby facilitating object-level alignment.

et al., 2025; Guo et al., 2025). Specifically, this architecture employs an MLP-based projector to align image features from a pretrained vision encoder with the input embedding space of an LLM backbone. These mapped visual tokens serve as soft prompts for conditional text generation. The training typically follows a two-stage strategy: where the projector is first pretrained on image-caption pairs before unfreezing the LLM for SFT (Liu et al., 2024a; Ye et al., 2023). The prevailing consensus in recent research is to utilize dense image captions during the modality alignment stage to provide detail-rich linguistic signals (Chen et al., 2024b; Li et al., 2024; Deitke et al., 2025).

Despite these advancements, models trained on holistic detailed captions suffer from a referential ambiguity issue: they must implicitly infer the correspondence between textual entities and specific visual regions from global representations. Consequently, models struggle to disentangle specific visual concepts from complex scenes, relying instead on statistical co-occurrences. This results in data inefficiency and suboptimal semantic grounding. In contrast, our approach seeks to provide explicit object-entity mappings, forcing the model to ground textual descriptions to their corresponding visual regions.

Fine-Grained Multimodal Alignment The limitations of image-level alignment have prompted increasing interest in fine-grained methodologies. For instance, SEA (Yin et al., 2025) utilizes CLIP (Radford et al., 2021) to assign semantic labels to visual patches and introduces a contrastive loss for patch-level alignment. Similarly, Patch Aligned Training (Jiang et al., 2025) employs off-the-shelf vision expert models (Zhang et al., 2024; Liu et al., 2024c; Kirillov et al., 2023) to generate text labels for visual patches and directly maximizes the cosine similarity between features of these visual patches and corresponding text embeddings.

While patch-level methods demonstrate improvements over image-level alignment, they overlook a fundamental semantic mismatch: single visual patches often lack independent semantic completeness, whereas a word carries distinct meaning. This fragmentation results in noisy alignment signals. In contrast, our method aligns text with complete visual objects rather than fragmented patches. By focusing on objects as the fundamental unit of visual semantics, our method ensures a cleaner, more semantically coherent alignment signal.

3 Methods

3.1 Insight: Code-Switching

Our method draws inspiration from the linguistic phenomenon of code-switching, defined as the interleaving of two or more languages within a single utterance (Thara and Poornachandran, 2018). For example, in the sentence "The *Klavier* is a versatile keyboard instrument," the German term *Klavier* (piano) functions as a code-switched entity embedded within the English context.

Conceptually, code-switching implies that the switched term must be semantically compatible with its surrounding context. We leverage this principle by treating visual objects as distinct "codes" carrying specific semantic information. Just as a bilingual speaker selects appropriate words from either language, we substitute textual entities with their visual counterparts. This imposes a constraint requiring the model to resolve the visual representation to satisfy the semantic expectations of the sentence, thereby achieving explicit object-level grounding.

3.2 Multimodal Code-Switching Pretraining

Building on this insight, we introduce Multimodal Code-Switching (MMCS), a novel alignment paradigm that establishes correspondence between visual objects and textual entities. As illustrated in Figure 2, we construct interleaved code-switching image-text sequences by replacing the entity tokens with their corresponding visual objects. This substitution strategy creates a dependency chain: the generation of the subsequent textual context, as well as the reconstruction of the entity itself, is directly conditioned on the visual counterparts. Therefore, the model is compelled to ground entities to their corresponding visual regions, eliminating the need to infer such correspondences solely from global representations.

Formally, let $\mathbf{X}$ denote the original text tokens. Consider a textual entity segment $\mathbf{e} = [x^i, \ldots, x^{i+m}]$ starting at index i. Let $\mathbf{v}_{\text{object}}$ denote the extracted visual object, defined as the set of image tokens whose spatial locations intersect with the object's bounding box. The interleaved image-text sequence $\mathbf{X}_{\text{MMCS}}$ is constructed by replacing the textual entity with the extracted object tokens:

$$\mathbf{X}_{\text{MMCS}} = \text{Concat}(\mathbf{X}^{<i}, \mathbf{v}_{\text{object}}, \mathbf{X}^{>i+m}). \quad (1)$$

We employ a language modeling objective where the loss is computed exclusively on text tokens, treating the visual tokens as conditioning context for next-token prediction:

$$\mathcal{L}_{\text{LM}} = -\sum_{x^t_{\text{MMCS}} \in \text{Text}} \log p_\theta(x^t_{\text{MMCS}} \mid \mathbf{X}^{<t}_{\text{MMCS}}). \quad (2)$$

To further enforce explicit supervision on object-entity correspondence, we additionally minimize the negative log-likelihood of the original textual entity segment given the preceding context:

$$\mathcal{L}_{\text{entity}} = -\log p_\theta(\mathbf{e} \mid \mathbf{X}^{<i}_{\text{MMCS}}, \mathbf{v}_{\text{object}}), \quad (3)$$

where $\mathbf{e}$ represents the substituted textual entity tokens in the original text sequence. The overall MMCS pretraining objective combines the language modeling loss and the entity reconstruction loss:

$$\mathcal{L}_{\text{MMCS}} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{entity}}. \quad (4)$$

Since visual tokens replace the textual entity e in X_{MMCS}, these two losses are non-overlapping and complementary. $\mathcal{L}_{\text{entity}}$ serves as the dedicated semantic anchor for entity reconstruction, while $\mathcal{L}_{\text{LM}}$ handles contextual integration.

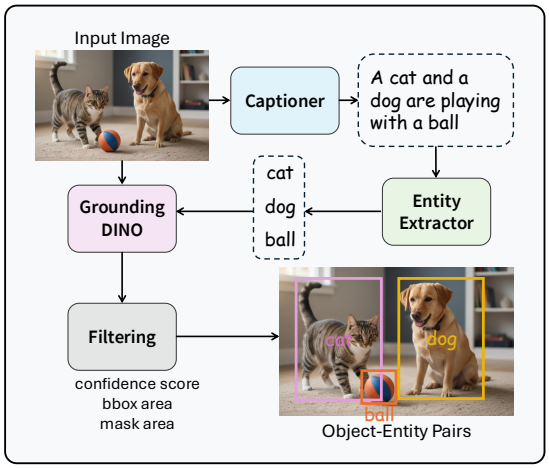


Figure 3: Data synthesis pipeline.

3.3 Data Synthesis Pipeline

As shown in Figure 3, we develop a data synthesis pipeline to generate interleaved code-switching image-text sequences with precise object-entity correspondence. For each input image, the pipeline proceeds through the following steps:

Detailed Image Captioning We begin by utilizing Qwen3-VL-32B-Instruct (Bai et al., 2025a) to generate detailed image captions. To capture granular visual details, we prompt the caption model to exhaustively describe the attributes of all visible elements within the scene.

Textual Entity Extraction Given the generated detailed captions, we employ Qwen2.5-72B-Instruct (Yang et al., 2024) to extract textual entities. The output comprises a list of noun phrases, often accompanied by brief attribute descriptions (e.g., a white mug labeled "O.CO"). These attributes are essential for disambiguating instances where multiple objects of the same category are present in a single image.

Visual Object Grounding We leverage Grounding DINO (Liu et al., 2024c) to anchor these textual entities to their corresponding visual regions. For each successfully localized entity, the output is a tuple containing the bounding box coordinates, the text label, and an associated confidence score.

Filtering To ensure high-quality, one-to-one object-entity correspondence, we implement a multi-stage filtering process. First, bounding boxes with low confidence scores are discarded, and redundant overlapping boxes are removed via Non-Maximum Suppression (NMS) (Neubeck and Gool,

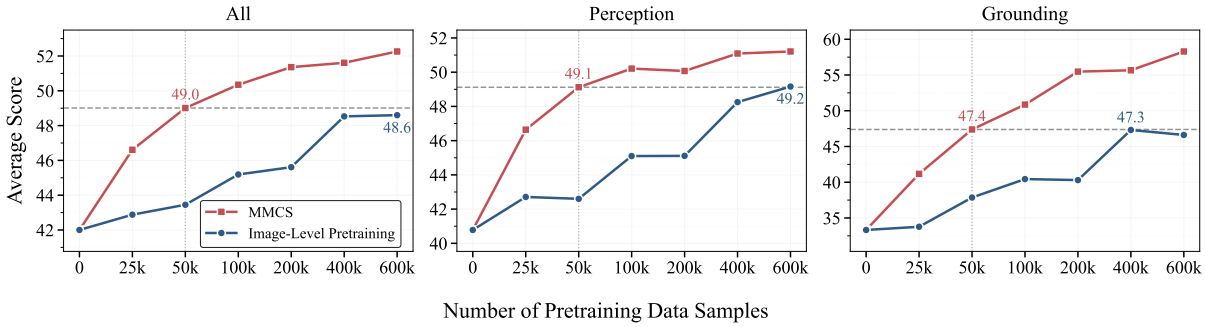


Figure 4: **Analysis of data efficiency.** We utilize Qwen2.5-3B as the LLM backbone and evaluate data efficiency by varying the pretraining dataset size from 0 to 600K, while keeping the SFT dataset fixed at 200K. The curves depict average scores across our benchmark suite, demonstrating that MMCS achieves superior data efficiency over image-level pretraining. Detailed compositions of the benchmarks used to calculate the average scores are provided in Appendix B.2.

2006). Subsequently, we calculate the area of each remaining bounding box and prune those violating spatial constraints (e.g., smaller than a single visual patch or exceeding 50% of the total image area). Finally, SAM-2.1 (Ravi et al., 2025) is employed to generate segmentation masks. Objects yielding mask areas below a predefined threshold are discarded to avoid the inclusion of heavily occluded instances.

We curate a diverse collection of images and apply the aforementioned synthesis pipeline. The generated pretraining dataset comprises 773K samples, each containing a source image, its detailed caption, and a set of visual objects paired with their bounding boxes and textual entity labels. A statistical summary and a detailed quality analysis of the synthesized dataset are presented in Appendix A.

4 Experiments

4.1 Experiment Setup

Model Architecture 1) **Vision Encoder**: Unless otherwise specified, we use SigLIP2-SO-400M (Tschannen et al., 2025) equipped with a tile-wise dynamic resolution strategy (Liu et al., 2024b). 2) **LLM**: MMCS is integrated with Qwen2.5-3B-Instruct (Yang et al., 2024), Qwen3-8B (Yang et al., 2025), and Llama3-8B-Instruct (Team, 2024). 3) **Projector**: We employ a two-layer MLP with a 2×2 pixel unshuffle operation (Chen et al., 2024d) to reduce the number of image tokens.

Training Dataset For pretraining, we use the 773K dataset synthesized via the pipeline described in Section 3.3 for both MMCS and image-level pretraining. For SFT, we adopt the LLaVA-NeXT (Liu et al., 2024b) dataset, which contains 779K high-quality instruction-following samples. We revert to using standard global image representations during SFT.

Evaluation Our evaluation covers a comprehensive suite of benchmarks categorized into three domains: 1) **Visual Grounding**: evaluating region understanding and grounding capabilities on referring expression comprehension (REC) tasks. 2) **Visual Perception**: including OCR-related tasks and real-world fine-grained perception. 3) **General VQA**: covering broad visual understanding and reasoning tasks.

For details on training setting and more evaluation results, please refer to Appendix B.

4.2 Analysis of Data Efficiency

We first investigate the data efficiency of MMCS by varying the scale of the pretraining dataset. Specifically, we compare MMCS against standard image-level supervision across dataset sizes ranging from 0 to 600K samples. To isolate the impact of pretraining, we use a constant subset of 200K examples from LLaVA-NeXT dataset for SFT. As illustrated in Figure 4, MMCS exhibits superior data efficiency. With only 50K samples, MMCS achieves downstream performance that surpasses image-level baseline trained on 600K samples. We attribute this advantage to the explicit object-entity supervision introduced in our method, which enables the model to efficiently align visual objects with their corresponding textual entities instead of relying on massive-scale statistical co-occurrences.

4.3 Scaling Up Training Data and Model Sizes

To investigate the universality of MMCS, we scale up the training to utilize the full 773K pretraining

LLM	Method	RefCOCO			RefCOCO+			RefCOCOg		AVG
		testA	testB	val	testA	testB	val	test	val	
Qwen2.5 3B	Caption	65.00	54.64	61.68	63.88	47.49	56.62	61.27	62.39	59.12
	MMCS	**80.66**	**68.75**	**75.18**	**74.35**	**56.23**	**65.49**	**70.32**	**71.31**	**70.29**
Qwen3 8B	Caption	**85.40**	74.15	81.51	80.89	64.98	72.82	76.61	78.76	76.89
	MMCS	85.02	**77.80**	**84.21**	**83.93**	**68.71**	**76.98**	**80.11**	**81.58**	**79.79**
Llama3 8B	Caption	73.29	56.23	65.94	70.10	50.62	60.59	63.86	64.65	63.16
	MMCS	**83.08**	**70.99**	**78.79**	**77.35**	**57.84**	**68.51**	**72.38**	**74.30**	**72.91**

Table 1: Performance on **referring expression comprehension**. "Caption" denotes image-level pretraining with image-caption pairs. We report Acc@0.5 across all splits.

LLM	ViT	Method	Perception				General		
			CVBench	OCR	VQA^T	V*	MMB	MME	MMStar
Qwen2.5 3B	SigLIP2	Caption	61.26	44.70	58.29	48.17	67.53	60.36	42.77
		MMCS	**66.87**	**47.60**	**59.35**	**50.79**	**68.56**	**62.75**	**43.98**
Qwen2.5 3B	QwenViT	Caption	56.33	58.80	62.87	57.59	67.70	62.82	45.21
		MMCS	**59.78**	**62.40**	**64.97**	57.59	**68.47**	**65.68**	**45.31**
Qwen3 8B	SigLIP2	Caption	69.71	50.60	64.37	50.79	74.48	68.14	48.67
		MMCS	**72.82**	**53.30**	**65.23**	**53.40**	**76.63**	**68.54**	**51.49**
Llama3 8B	SigLIP2	Caption	65.66	47.00	62.59	51.31	**70.79**	63.43	**44.96**
		MMCS	**69.48**	**48.90**	**64.50**	**53.93**	68.99	**65.54**	43.15

Table 2: Performance on **perception-centric** and **general VQA** benchmarks. "OCR", "VQA^T", "V*", and "MMB" denote OCRBench, TextVQA, V-Star, and MMBench, respectively. We report the normalized score for MME.

and 779K instruction-following datasets. To verify robustness across different model sizes and architectures, we extend our evaluation to larger LLMs, including Qwen3-8B and Llama3-8B.

Visual Grounding We evaluate visual grounding capabilities on referring expression comprehension tasks. As shown in Table 1, MMCS exhibits considerable improvements over the image captioning baseline across RefCOCO/+/g datasets, achieving an average gain of 7.9%. Notably, since MMCS does not introduce bounding box coordinates of visual objects during pretraining, these improvements stem directly from enhanced object recognition and region understanding capabilities, highlighting the effectiveness of explicit object-level supervision during modality alignment.

Visual Perception As shown in Table 2, MMCS consistently outperforms standard image captioning paradigm in perception tasks, maintaining robust gains as data volumes and model sizes increase. Our method demonstrates considerable performance gains in fine-grained visual perception tasks, achieving average improvements of 4.0% on CVBench, 2.8% on OCRBench and 2.0% on V-Star. These results suggest that by grounding textual descriptions to specific visual regions, MMCS effectively facilitates precise visual perception.

General VQA MMCS yields improvements on general VQA tasks as well, as shown in Table 2. We attribute these gains to a dual mechanism: first, enhanced visual perception and region understanding provide a more accurate visual context; second, the high-quality modality alignment establishes a stronger foundation for obtaining multimodal reasoning capability in the SFT stage.

For more evaluation results, please refer to Appendix B.2.

4.4 Generalization Across Vision Encoders

Beyond tile-wise approaches, native dynamic resolution (Wang et al., 2024; Bai et al., 2025b) constitutes another prominent strategy for processing high-resolution images. This strategy encodes images into a variable number of visual tokens while preserving original aspect ratios (Dehghani et al., 2023). We demonstrate our method's general effectiveness using QwenViT, the vision encoder utilized in Qwen2.5-VL (Bai et al., 2025b) which employs this native dynamic resolution mechanism. As evidenced in Table 2, MMCS maintains its superior performance in this setting. These results

Method	General	Perception	Grounding
Caption	51.19	55.77	59.12
SEA	50.78	55.69	60.96
Patch Aligned	51.78	56.43	66.55
MMCS (ours)	**51.84**	**58.24**	**70.29**
w/o $\mathcal{L}_{\text{entity}}$	51.45	54.23	67.55
w/o $\mathcal{L}_{\text{LM}}$	51.37	56.26	65.44

Table 3: Performance comparison with **patch-level alignment** and **ablation study**. All experiments use Qwen2.5-3B-Instruct as the LLM backbone. See Appendix B.2 for the benchmark compositions of these average scores.

validate our method's compatibility across diverse vision encoders.

4.5 Comparison with Patch-Level Alignment

To ensure a fair comparison with patch-level alignment methods, we reproduce SEA (Yin et al., 2025) and Patch Aligned Training (Jiang et al., 2025) using the same pretraining and SFT datasets as MMCS. As presented in Table 3, MMCS consistently outperforms patch-level baselines across all downstream tasks. These results confirm that mapping textual entities to distinct visual objects rather than arbitrary patches establishes a more robust and semantically coherent alignment. Full results are available in Appendix B.2.

4.6 Ablation Study

We perform an ablation study on the components of Eq. 4. The results in Table 3 demonstrate that both objectives are crucial for optimal performance.

Role of $\mathcal{L}_{\text{entity}}$: Enforcing Semantic Precision. Ablating the entity reconstruction loss ($\mathcal{L}_{\text{entity}}$) results in the most severe drop in visual perception performance (-4.0%). This indicates that $\mathcal{L}_{\text{entity}}$ acts as a *semantic anchor*. By forcing the model to explicitly translate visual features back into textual entities, this objective compels the projector to encode precise and discriminative visual details, thereby enhancing visual perception.

Role of $\mathcal{L}_{\text{LM}}$: Facilitating Contextual Integration. Conversely, removing the language modeling loss ($\mathcal{L}_{\text{LM}}$) primarily degrades visual grounding capabilities (-4.9%). We attribute this to $\mathcal{L}_{\text{LM}}$'s role in *contextual integration*. It encourages the model to reason about the relationship between the visual object and its descriptive attributes, which is critical for both visual perception and referring expression comprehension.

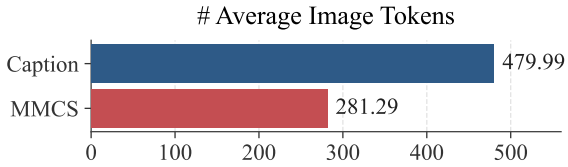


Figure 5: Comparison of the average number of image tokens between MMCS and standard image-level pretraining.

5 Further Discussion

5.1 Analysis of Training Cost

In standard image-level pretraining, the entire image is encoded into a fixed set of image tokens. In contrast, MMCS operates explicitly on localized visual objects. As defined in Eq. 1, only the image patches that spatially intersect with the target bounding boxes are retained and fed into the LLM backbone, inherently reducing the input sequence length. To quantify this, we randomly selected 3,000 instances from each source dataset and compared the average number of image tokens per sample under identical dynamic resolution configurations. As shown in Figure 5, MMCS reduces the token count by approximately 41.4% compared to standard image-level pretraining.

5.2 Representation Alignment Measurement

To investigate whether MMCS achieves superior multimodal alignment at the feature level compared to standard image-level pretraining, we directly quantify vision-language representational alignment after pretraining. We employ three metrics: CKA (Kornblith et al., 2019) to evaluate *global* geometric correspondence, and Mutual k-NN and CKNNA (Huh et al., 2024) to assess *local* neighborhood consistency. Detailed definitions of these metrics are provided in Appendix D.

Specifically, we input 1,000 image-caption pairs sampled from the COCO2014 validation set into the MLLM and extract hidden states from all layers. These states are subsequently partitioned into visual and textual components to compute layer-wise alignment metrics. As shown in Figure 6, MMCS achieves superior representational alignment compared to image-level supervision. These findings align with the hypothesis that representational alignment correlates with model capability (Huh et al., 2024), offering a rationale for the performance improvements observed in downstream tasks.

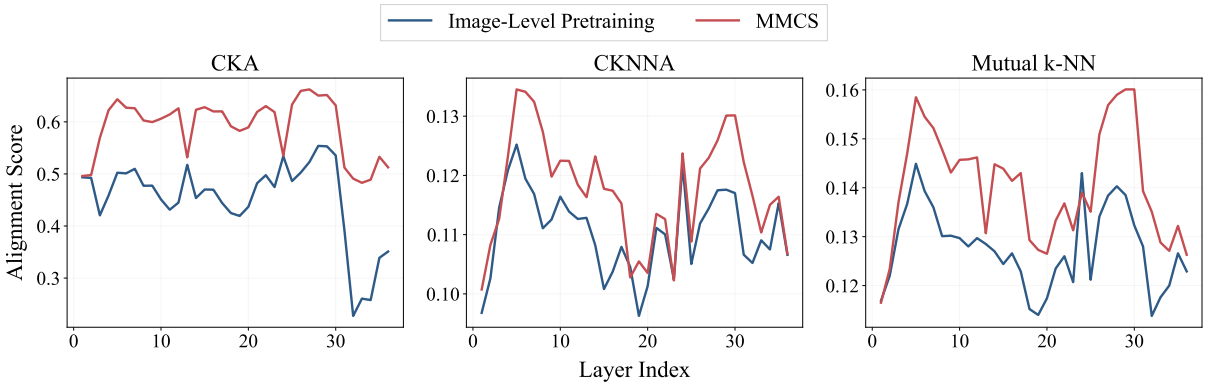


Figure 6: **Layer-wise representation alignment.** To evaluate cross-modal integration, we extract hidden states from each LLM decoder layer for input image-text pairs and partition them by modality. We then quantify the alignment using three distinct metrics: CKA, CKNNA and Mutual k-NN. MMCS exhibits superior representation alignment across the majority of layers over the standard image-level pretraining.

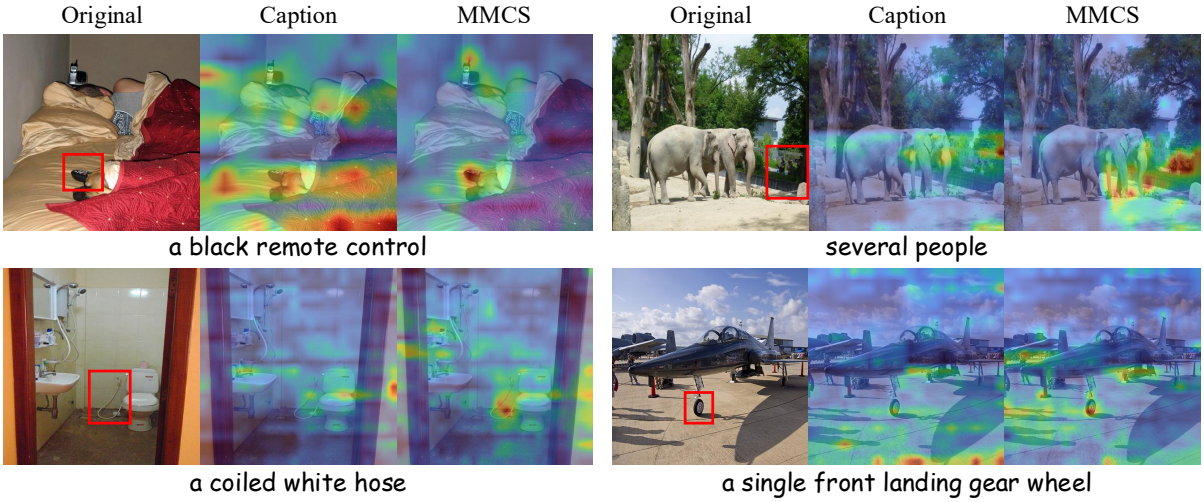


Figure 7: **Visualization of attention maps.** Each example displays a triplet containing: (Left) the original image with a red bounding box highlighting the target object; (Middle) the attention map from an MLLM pretrained with standard captioning; and (Right) the attention map from our MMCS method. Warmer colors indicate higher attention weights (min-max normalized). The specific text entity driving the attention is noted below each example.

5.3 Attention Map Analysis

Figure 7 presents a qualitative comparison of attention maps. We analyze the cross-modal attention distribution from textual entity tokens to visual features by feeding image-caption pairs into the MLLM. As observed, the model pretrained with MMCS accurately attends to visual regions corresponding to specific textual entity descriptions. In contrast, the model trained with image-caption pairs often produces diffuse or misaligned attention patterns. Given that the LLM backbone is kept frozen during pretraining, these results indicate that training the projector with explicit object-entity correspondence induces the learning of more semantically precise and interpretable features. These features align better with the LLM's pre-existing semantic space. More qualitative results are presented in Appendix E.

6 Conclusion

In this paper, we propose MMCS, a modality alignment pretraining paradigm that establishes explicit object-entity correspondence. Our extensive experiments reveal that MMCS demonstrates considerable improvements in both data efficiency and downstream performance. Furthermore, we investigate the advantages of MMCS in facilitating multimodal representation alignment from the perspective of the internal feature space topology. Our findings highlight the importance of object-level alignment in developing data-efficient MLLMs with advanced performance.

Limitations

While MMCS demonstrates considerable improvements on data efficiency and downstream performance, our current implementation primarily focuses on visual objects within natural images. The core methodology of establishing explicit correspondence between local visual features and textual entities is generalizable to broader scenarios. Extending MMCS to domains such as chart understanding or scene text recognition will be further explored in our future works. Furthermore, exploring bootstrapping mechanisms for iterative self-improvement and extending the MMCS paradigm to encompass complex relations and actions represent highly promising directions that we intend to investigate.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *CoRR*, abs/2509.23661.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for A lite vision-language model. *CoRR*, abs/2402.11684.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024c. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Google DeepMind. 2025. Gemini 3.1 pro. `https://deepmind.google/models/gemini/pro/`.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M. Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey A. Gritsenko, Mario Lucic, and Neil Houlsby. 2023. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA,*

June 11-15, 2025, pages 91–104. Computer Vision Foundation / IEEE.

Hongyuan Dong, Zijian Kang, Weijie Yin, LiangXiao LiangXiao, ChaoFeng ChaoFeng, and Ran Jiao. 2025. Scalable vision language model training via high quality data curation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 33272–33293. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 93–127. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 81 others. 2025. Seed1.5-vl technical report. *CoRR*, abs/2505.07062.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. 2025. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *CoRR*, abs/2505.17316.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5648–5656. Computer Vision Foundation / IEEE Computer Society.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 235–251. Springer.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab

Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981.

Xiangtai Li, Tao Zhang, Yanwei Li, Haobo Yuan, Shihao Chen, Yikang Zhou, Jiahao Meng, Yueyi Sun, Shilin Xu, Lu Qi, Tianheng Cheng, Yi Lin, Zilong Huang, Wenhao Huang, Jiashi Feng, and Guang Shi. 2025. Denseworld-1m: Towards detailed dense grounded caption in the real world. *CoRR*, abs/2506.24102.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. 2024. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024c. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pages 38–55. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024e. Ocrbench: on the hidden mystery of OCR in large multimodal models. *Sci. China Inf. Sci.*, 67(12).

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, and 10 others. 2024. MM1: methods, analysis and insights from multimodal LLM pre-training. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, volume 15087 of *Lecture Notes in Computer Science*, pages 304–323. Springer.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.

Alexander Neubeck and Luc Van Gool. 2006. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, pages 850–855. IEEE Computer Society.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. DOCCI: descriptions of connected and contrasting images. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024,*

Proceedings, Part LX, volume 15118 of *Lecture Notes in Computer Science*, pages 291–309. Springer.

Shana Poplack. 1981. Syntactic structure and social function of code-switching. In Richard P. Durán, editor, *Latino Language and Communicative Behavior*, pages 169–184. Ablex, Norwood, NJ.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2025. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8429–8438. IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

S. Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, Bangalore, India, September 19-22, 2018*, pages 2382–2388. IEEE.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13084–13094. IEEE.

xAI Team. 2024. Grok-1.5 Vision Preview.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. 2025. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *CoRR*, abs/2509.22647.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Yuanxing Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Wentao Zhang, and Feng Zhao. 2025. SEA: Supervised embedding alignment for token-level visual-textual integration in MLLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1058–1070, Suzhou, China. Association for Computational Linguistics.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Xiao-Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, and 54 others. 2025. Mimo-vl technical report. *CoRR*, abs/2506.03569.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 881–916. Association for Computational Linguistics.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. 2024. Recognize anything: A strong image tagging model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 1724–1732. IEEE.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479.

A Training Dataset Details

A.1 MMCS Pretraining Dataset

# Samples	# Objects	Avg. Chars	Avg. Objects
773,779	5,145,630	961.65	6.65

Table 4: **Statistics of the synthesized pretraining dataset.** "Avg. Chars" denotes the average character count per detailed caption, while "Avg. Objects" represents the average number of visual objects identified per sample after filtering.

The image sources utilized in Section 3.3 include MS COCO (Lin et al., 2014), Flickr30k (Young et al., 2014), GQA (Hudson and Manning, 2019), Objects365 (Shao et al., 2019), OpenImages (Kuznetsova et al., 2020), and SA-1B (Kirillov et al., 2023). We select 773K images from these sources and apply our data synthesis pipeline to construct the pretraining dataset. Table 4 presents a statistical summary of the dataset, and Figure 8 provides two illustrative examples.

Prompts for VLM Judge

You are judging whether a bounding box correctly localizes the target described by a phrase. The image contains a highlighted bbox. Decide if it covers the intended object.

Target phrase: {phrase}
Bbox (xyxy): [{x1, y1, x2, y2}]
Image size (w×h): [{width, height}]

Criteria:
- YES: the highlighted region matches the phrase, and the target object is mostly inside the box.
- NO: the box highlights the wrong object or attribute, or the target is clearly outside the box.

Answer exactly: YES or NO

Data Quality To validate the quality of our synthesized pretraining dataset, we conduct quantitative evaluations using both Gemini-3-Pro (DeepMind, 2025) as an automated Vision-Language

Judge	COCO	Flickr30k	GQA	Objects365	OpenImages	SA-1B	AVG
VLM	0.935	0.920	0.965	0.980	0.970	0.925	0.949
Human	0.950	0.900	0.900	1.000	1.000	0.900	0.942

Table 5: Quality evaluation of the synthesized MMCS pretraining dataset. Both the VLM judge and human annotators assign binary scores indicating whether a textual entity is accurately grounded by its corresponding visual bounding box. The reported values represent the average grounding accuracy.

Model (VLM) judge, alongside human verification. For each source dataset, we sample 200 objects to compute the VLM score. Specifically, the judge is prompted to assign a binary score (1 for correct, 0 for incorrect) indicating whether a textual entity is accurately localized by the provided bounding box. To assess the reliability of the VLM evaluations, human annotators independently verify a subset of 20 randomly selected objects per dataset. The evaluation results, representing the average grounding accuracy, are summarized in Table 5.

Instruction for Human Annotators

Task:
Each sample shows an image with a colored bounding box and a target phrase. Judge whether the box correctly highlights the object described by the phrase.

Label: YES or NO

YES: acceptable bbox
- The box covers the correct object matching the phrase.
- The object is mostly inside the box (minor over-/under-coverage is fine).
- Slight attribute ambiguity is tolerable if the object identity is correct. e.g. phrase="red car", the car is dark red — still YES.

NO: wrong bbox
- Wrong object: the box highlights a different object than what the phrase describes.
- Wrong attribute: any key attribute clearly conflicts.
- Missed target: the described object is mostly outside the box.

A.2 SFT Dataset

For SFT, we utilize the LLaVA-NeXT dataset (Liu et al., 2024b), which comprises 779K instruction-following samples covering general VQA, OCR-related tasks and document/chart understanding. Specifically, this dataset incorporates data from AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), DVQA (Kafle et al., 2018), GQA, LAION-GPT4V (Schuhmann et al., 2022), OCR-VQA (Mishra et al., 2019), ShareGPT4V (Chen et al., 2024b), SynthDoG-EN (Kim et al., 2022) and Visual Genome (Krishna et al., 2017).

Hyperparameter	Pretrain	Finetune
Trainable	MLP	MLP+LLM
Global batch size	128	64
Learning rate	1e-3	2e-4
Epochs	1	
Max image tiles	4 + 1	
Optimizer	AdamW	
LR schedule	Cosine decay	
Warm up ratio	0.03	
Weight decay	0	
LoRA r	-	32
LoRA alpha	-	64
LoRA dropout	-	0.05

Table 6: **Hyperparameters for model training.** "4+1" indicates that the high-resolution image is divided into at most 4 localized tiles with an additional thumbnail tile that provides a global view.

B Implementation Details

B.1 Training Hyperparameters

Table 6 details the hyperparameters employed for both modality alignment pretraining and SFT stages. High-resolution images are divided into smaller image tiles of the resolution that the ViT is originally trained for, and encoded independently. All experiments are conducted on 8 NVIDIA A6000 GPUs. During the SFT stage, we apply LoRA (Hu et al., 2022) to the LLM backbone. For 8B-scale models, pretraining completes within 10 hours, while SFT completes within 20 hours.

LLM	Method	RefCOCO			RefCOCO+			RefCOCOg		AVG
		testA	testB	val	testA	testB	val	test	val	
Qwen2.5 3B	Caption	65.00	54.64	61.68	63.88	47.49	56.62	61.27	62.39	59.12
	SEA	68.33	60.46	64.08	62.81	48.02	57.80	61.09	65.05	60.96
	Patch Aligned	77.82	64.08	71.84	71.34	50.83	61.95	66.67	67.90	66.55
	MMCS	**80.66**	**68.75**	**75.18**	**74.35**	**56.23**	**65.49**	**70.32**	**71.31**	**70.29**

Table 7: Full results on **referring expression comprehension**. We report Acc@0.5 across all splits.

LLM	ViT	Method	AI2D	ChartQA	CVBench	OCR	RWQA	VQA^T	V*	Avg
Qwen2.5 3B	SigLIP2	Caption	66.13	56.52	61.26	44.70	55.29	58.29	48.17	55.77
		SEA	66.77	53.48	63.15	44.90	55.42	58.48	47.64	55.69
		Patch Aligned	67.45	56.08	61.87	46.70	55.16	**59.58**	48.17	56.43
		MMCS	**68.72**	**57.36**	**66.87**	**47.60**	**56.99**	59.35	**50.79**	**58.24**
Qwen2.5 3B	QwenViT	Caption	69.40	70.36	56.33	58.80	55.16	62.87	57.59	61.50
		MMCS	**71.15**	**71.92**	**59.78**	**62.40**	**56.86**	**64.97**	57.59	**63.52**
Qwen3 8B	SigLIP2	Caption	72.51	62.48	69.71	50.60	58.17	64.37	50.79	61.23
		MMCS	**73.93**	**64.96**	**72.82**	**53.30**	**60.13**	**65.23**	**53.40**	**63.40**
Llama3 8B	SigLIP2	Caption	**72.31**	58.28	65.66	47.00	**59.87**	62.59	51.31	59.57
		MMCS	71.96	**60.20**	**69.48**	**48.90**	59.08	**64.50**	**53.93**	**61.15**

Table 8: Full results on **perception-centric** benchmarks. "OCR", "RWQA", "VQA^T", and "V*" denote OCRBench, RealWorldQA, TextVQA, and V-Star, respectively.

B.2 Benchmarks

We conduct a comprehensive evaluation across a suite of benchmarks organized into three distinct domains:

- **Visual Grounding**: We evaluate referring expression comprehension capability using RefCOCO, RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016).
- **Perception-centric tasks**: We employ AI2D and ChartQA for diagram and chart understanding; OCRBench (Liu et al., 2024e) and TextVQA (Singh et al., 2019) to assess OCR capabilities; and CVBench (Tong et al., 2024), RealWorldQA (xAI Team, 2024), and V-Star (Wu and Xie, 2024) for real-world visual perception.
- **General VQA**: We encompass MMBench (Liu et al., 2024d), MME (Fu et al., 2023), MMStar (Chen et al., 2024c), and MMVet (Yu et al., 2024) for general visual understanding; MMMU (Yue et al., 2024) for multidisciplinary reasoning requiring college-level knowledge; and GQA for real-world compositional visual reasoning.

To ensure reproducibility, all evaluations are conducted using the lmms-eval framework (Zhang et al., 2025). The full evaluation results are in Table 7, Table 8, and Table 9.

C Additional Experimental Analysis

In this section, we present further experimental analyses to evaluate the scalability and robustness of the proposed MMCS paradigm. To rigorously isolate the impact of variations in the pretraining stage and ensure fair comparisons across all settings, all models evaluated in this section undergo an identical SFT phase. Specifically, consistent with the data efficiency analysis in Section 4.2, we keep the SFT data fixed to a subset of 200K samples from the LLaVA-NeXT dataset.

C.1 Impact of Image Resolution

We evaluate the robustness of MMCS across varying input image resolutions. By adjusting the maximum number of tiles (ranging from 1 to 12), we compare our method against the standard caption baseline with a fixed set of 600K pretraining samples. As shown in Table 10, higher dynamic resolutions generally yield better performance in perception and grounding tasks. Crucially, MMCS consistently outperforms the baseline across all resolution configurations, underscoring the efficacy of our method regardless of the visual input granularity.

LLM	ViT	Method	MMB	MME	MMStar	MMMU	MMVet	GQA	Avg
Qwen2.5 3B	SigLIP2	Caption	67.53	60.36	40.60	42.77	**35.64**	60.22	51.19
		SEA	68.81	60.86	40.40	42.54	31.83	60.26	50.78
		Patch Aligned	**69.76**	**63.71**	**41.30**	43.45	32.11	60.36	51.78
		MMCS	68.56	62.75	40.90	**43.98**	33.30	**61.57**	**51.84**
Qwen2.5 3B	QwenViT	Caption	67.70	62.82	39.20	45.21	39.27	60.12	52.39
		MMCS	**68.47**	**65.68**	**42.90**	**45.31**	**39.31**	**60.66**	**53.72**
Qwen3 8B	SigLIP2	Caption	74.48	68.14	47.10	48.67	**44.72**	62.54	57.61
		MMCS	**76.63**	**68.54**	**48.80**	**51.49**	41.56	**63.02**	**58.34**
Llama3 8B	SigLIP2	Caption	**70.79**	63.43	39.20	**44.96**	39.36	63.17	53.48
		MMCS	68.99	**65.54**	**40.90**	43.15	**42.06**	**63.40**	**54.01**

Table 9: Full results on **general VQA** benchmarks. We report the normalized score for MME, and evaluate MMVet using GPT-4.1 as the judge.

C.2 Experiments with more data

To investigate the scalability of our approach, we further expand the pretraining dataset from 773K to 1M samples by incorporating additional images from Objects365 and SA-1B. All newly added samples are processed using the identical data synthesis pipeline detailed in Section 3.3. As presented in Table 11, MMCS maintains an advantage over standard image-level pretraining across all data scales. This confirms that our explicit object-level alignment paradigm can effectively leverage larger volumes of pretraining data to achieve more robust semantic grounding.

C.3 Analysis of OCR Capability Enhancements

As shown in Table 2, MMCS improves performance on OCR-related benchmarks (e.g., OCRBench, TextVQA) despite the absence of dedicated OCR pretraining data. We attribute this emergent capability to two primary factors:

Implicit OCR Supervision from Natural Objects Natural images frequently contain objects with embedded text (e.g., street signs, product labels). While standard image-level pretraining heavily compresses these small regions into global features, MMCS explicitly crops and aligns them with textual descriptions, acting as an implicit OCR training objective. To quantify this, we utilized Gemini-3-Pro to evaluate 500 random object crops per source dataset. As shown in Table 12, a notable proportion of natural objects (10.0% to 30.8%) contain legible text, providing substantial implicit supervision.

Generalization via Enhanced Modality Alignment By explicitly resolving the referential ambiguity inherent in global image-text pairs, MMCS establishes a more precise modality alignment. This fundamental improvement enables the model to accurately localize and parse fine-grained visual details, a capability that naturally generalizes to text recognition tasks.

Max Res.	Method	General	Perception	Grounding
$384^2 \times 1$	Caption	46.62	38.32	41.78
	MMCS	49.45	42.75	44.76
$384^2 \times 4$	Caption	48.95	49.16	46.61
	MMCS	50.47	51.21	58.29
$384^2 \times 9$	Caption	48.70	50.51	46.05
	MMCS	50.86	53.08	58.96
$384^2 \times 12$	Caption	49.00	48.36	50.44
	MMCS	50.73	54.39	58.07

Table 10: Performance comparison across **different maximum image resolutions**. The input resolution is dynamically scaled by varying the maximum number of 384×384 tiles. MMCS consistently outperforms the standard image-level pretraining baseline.

# Samples	Method	General	Perception	Grounding
50K	Caption	47.25	42.60	37.86
	MMCS	49.68	49.12	47.38
600K	Caption	48.95	49.16	46.61
	MMCS	50.47	51.21	58.29
1M	Caption	48.60	48.78	49.85
	MMCS	50.04	51.51	60.27

Table 11: Performance comparison across varying pretraining dataset sizes. MMCS maintains an advantage over the standard baseline at 1M pretraining scales.

D Representation Alignment Metrics

In this section, we provide detailed definitions of the representation alignment metrics used in Sec-

Metric	COCO	Flickr30k	GQA	Objects365	OpenImages	SA-1B
# Has Text	91	50	107	82	119	154
Ratio (%)	18.2	10.0	21.4	16.4	23.8	30.8

Table 12: Proportion of visual objects containing legible text across different source datasets. Statistics are based on a random sample of 500 objects per dataset, evaluated by a VLM judge.

tion 5.2. We adopt the notation from Huh et al. (2024).

CKA CKA (Centered Kernel Alignment) reflects the *global* similarity between two representation spaces by comparing their kernel matrices. Let $\phi_i \in \mathbb{R}^n$ and $\psi_i \in \mathbb{R}^m$ be vectorized features of two modalities (e.g. language and vision). Let $\mathbf{K}_{ij} = \kappa(\phi_i, \phi_j)$ and $\mathbf{L}_{ij} = \kappa(\psi_i, \psi_j)$ be the kernel matrices computed from a dataset using some kernel function κ. For an inner-product kernel, the ij-th entry of the centered counterpart of these kernel matrices is given by

$$\begin{aligned} \bar{\mathbf{K}}_{ij} &= \langle \phi_i, \phi_j \rangle - \mathbb{E}_l[\langle \phi_i, \phi_l \rangle], \\ \bar{\mathbf{L}}_{ij} &= \langle \psi_i, \psi_j \rangle - \mathbb{E}_l[\langle \psi_i, \psi_l \rangle]. \end{aligned} \tag{5}$$

Then the cross-covariance of $\mathbf{K}$ and $\mathbf{L}$ is:

$$\text{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} \text{Trace}(\bar{\mathbf{K}}\bar{\mathbf{L}}). \tag{6}$$

Finally, CKA is obtained by normalizing this quantity:

$$\text{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\text{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\text{HSIC}(\mathbf{K}, \mathbf{K})\text{HSIC}(\mathbf{L}, \mathbf{L})}}. \tag{7}$$

CKNNA CKNNA (Centered Kernel Nearest-Neighbor Alignment) is a relaxed variation of CKA that emphasizes *local* structural alignment. It modifies the measure by replacing $\text{HSIC}(\mathbf{K}, \mathbf{L})$ with $\text{Align}(\mathbf{K}, \mathbf{L})$, which computes Eq. 6 considering only the k-nearest neighbors in the dataset:

$$\text{Align}(\mathbf{K}, \mathbf{L}) = \sum_i \sum_j \alpha(i, j) \bar{\mathbf{K}}_{ij} \bar{\mathbf{L}}_{ij}, \tag{8}$$

where $\alpha(i, j)$ is an indicator function that selects common nearest neighbors:

$$\alpha(i, j) = \mathbb{1}[\phi_j \in \text{knn}(\phi_i) \wedge \psi_i \in \text{knn}(\psi_j) \wedge i \neq j]. \tag{9}$$

This term acts as a mask, preserving only the interactions between sample i and j if j is a neighbor of i in both representation spaces.

Mutual k-NN Mutual k-NN (Mutual k-Nearest Neighbor) measures the average overlap of nearest neighbor sets of representations. Let $\{\phi_i, \psi_i\}_{i=1}^b$ denote a mini-batch of paired features from two modalities, where the collections of these features are denoted as $\Phi = \{\phi_1, \ldots, \phi_b\}$ and $\Psi = \{\psi_1, \ldots, \psi_b\}$. For each feature pair (ϕ_i, ψ_i), we compute the respective nearest neighbor sets $\mathcal{S}(\phi_i)$ and $\mathcal{S}(\psi_i)$:

$$\begin{aligned} \mathcal{S}(\phi_i) &= d_{\text{knn}}(\phi_i, \Phi \setminus \phi_i), \\ \mathcal{S}(\psi_i) &= d_{\text{knn}}(\psi_i, \Psi \setminus \psi_i), \end{aligned} \tag{10}$$

where d_{knn} returns the set of indices of the k-nearest neighbors. We then measure the alignment via the average intersection:

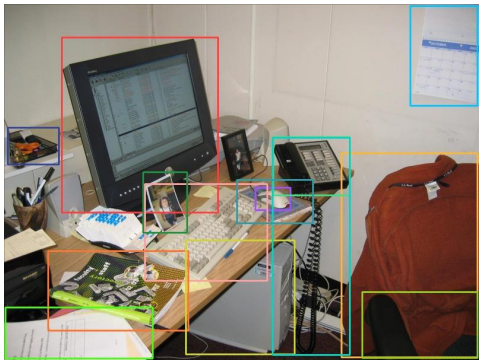
$$m_{\text{NN}}(\phi_i, \psi_i) = \frac{1}{k} |\mathcal{S}(\phi_i) \cap \mathcal{S}(\psi_i)|, \tag{11}$$

where $|\cdot|$ denotes the cardinality of the set.

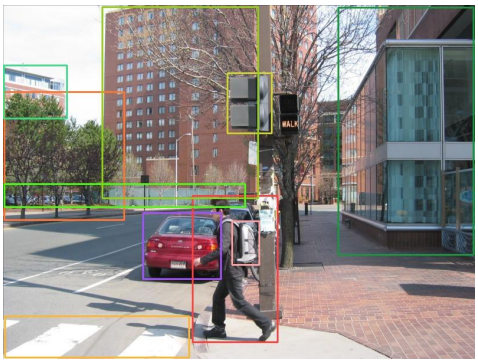
In Section 5.2, CKA is calculated using all 1000 image-caption pairs in the dataset. For CKNNA and Mutual k-NN, we report results with $k = 10$.

E More Qualitative Results

Figure 9 presents a representative case demonstrating our model's dynamic attention behavior when processing scenes with multiple target objects. Specifically, given an input image of a framed mosaic artwork alongside a dense caption, we observe clear spatial shifts in the model's visual attention as it processes different textual entities. This capability indicates that our model can effectively disentangle multiple localized objects within a single complex image. Furthermore, this dynamic localization behavior suggests that the explicit object-entity correspondence established by our MMCS paradigm successfully guides the model to decouple semantic regions.



A black CRT computer monitor is turned on, displaying a windowed interface with lists of text. In front of the monitor sits a beige keyboard with raised keys and a blue pen resting on it. To the right of the keyboard is a white computer mouse on a blue mousepad. A black landline telephone with a coiled cord is placed near the mouse. The desk surface is wooden and cluttered with various items including a green and black notebook with a circuit board design, yellow sticky notes, a stack of white papers, and a small photo frame showing a person. Another smaller photo frame with a picture of two people is positioned near the monitor. A white desktop computer tower is under the desk. On the left side of the desk, there is a container holding pens and markers, a paper clip holder, and a tray with scissors and other office supplies. A calendar for October is hanging on the wall to the right. An orange sweater or jacket is draped over the back of a black office chair. The wall behind the desk is plain and light-colored, with visible vertical paneling



A man wearing a dark jacket, black pants, and black shoes is walking across a crosswalk. He has short dark hair and is carrying a gray backpack. The crosswalk consists of white painted stripes on the asphalt road. A pedestrian signal mounted on a metal pole displays the word \"WALK\" in orange letters. Behind the man, a red compact car with a visible license plate is stopped at the curb. The sidewalk is made of red brick pavers and runs alongside a modern building with large glass windows featuring a blue-tinted pattern. Several bare trees line the sidewalk and street. In the background, a tall red-brick apartment or office building with many windows stands behind a green chain-link fence. To the left, there are green-leaved trees and another multi-story building with a blue facade. The sky is bright and mostly clear. A few other vehicles are parked along the street further back.

Figure 8: Illustrative examples from the pretraining dataset. **Left**: The original image with annotated bounding boxes of visual objects. **Right**: The dense caption of the image, in which successfully localized textual entities are highlighted. The textual entities are color-coded to match the corresponding grounded visual objects.

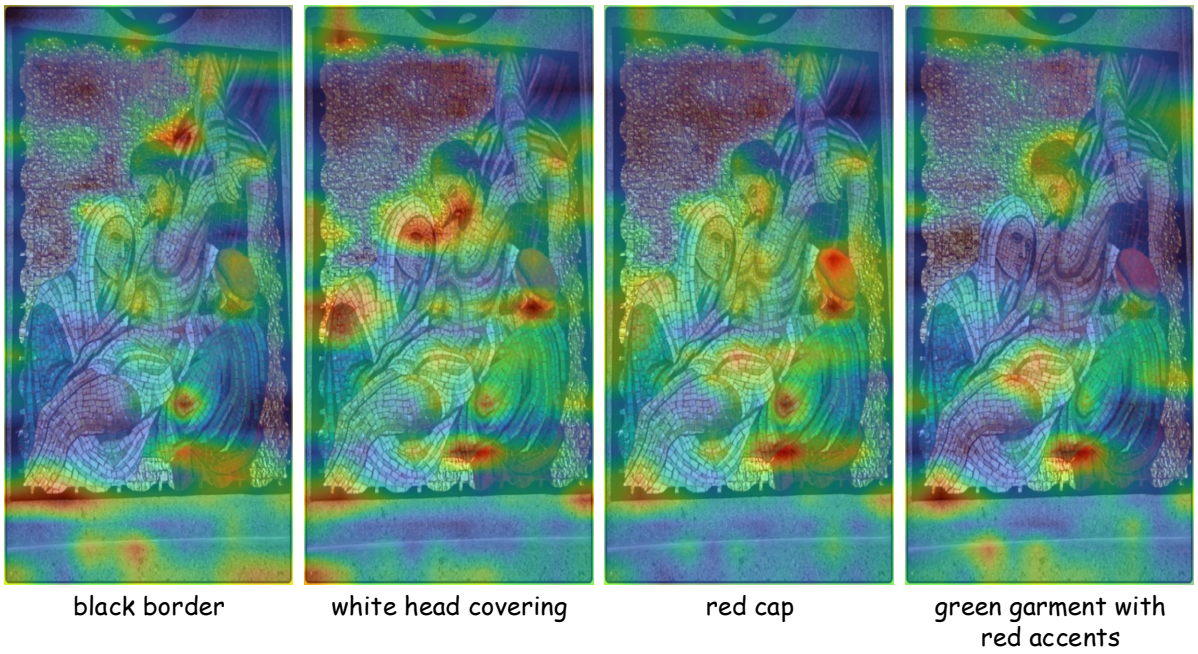


Figure 9: Visualization of dynamic attention shifts for a complex scene containing multiple target objects. The heatmaps illustrate how the model's visual attention shifts to focus on the specific region as the corresponding textual entity is processed.