Toward Compact and Structured Visual Representations in VLMs: SSM-Based Vision Encoders as an Alternative to Transformers

Anonymous CVPR submission

Paper ID *****

Abstract

Spatial concept understanding is important for vision-language models (VLMs), not only for explicit grounding and localization tasks but also for general visual question answering. The common solution to this challenge has been to scale image resolution or visual token counts, yet this conflicts with the quadratic attention cost of ViT-based encoders that nearly all current VLMs rely on. We explore an orthogonal question and investigate whether a vision encoder can produce compact and structured visual token representations that carry richer spatial concept information under a fixed token budget. We conduct the first controlled evaluation of SSM-based vision encoders as frozen visual backbones in VLMs and find that VMamba-based VLMs achieve stronger spatial concept understanding than ViT-family alternatives at matched scale and token budget, with improvements that transfer from grounding benchmarks to general visual question answering. Token-region similarity maps computed at intermediate layers of the language model show that SSM visual tokens produce sharper and more spatially selective concept-region binding during LLM reasoning, showing that the language model can use SSM representations better than ViT-based vision backbones under similar settings. These findings suggest that architectural inductive bias is an underexplored direction for improving visual concept representations in VLMs without increasing token budgets.

1. Introduction

Spatial concept understanding is a core requirement for vision-language models (VLMs). Beyond explicit grounding and localization tasks, spatial understanding also plays an important role in general visual question answering. For example, GQA [10] is built around scene graphs with explicit spatial relationships, and several works show that improving visual grounding leads to better performance on both GQA and VQA-v2 [14, 21].

A recurring challenge in VLMs is extracting spatially grounded evidence from images under a fixed token budget [4, 9]. The dominant response has been to increase image resolution or the number of visual tokens [15, 29]. However, this approach has a fundamental limitation: virtually all current VLMs rely on ViT-based vision encoders [12], whose self-attention mechanism scales quadratically with the number of tokens. Increasing resolution or token count therefore quickly becomes computationally infeasible. More importantly, this approach treats a representation quality problem as a quantity problem, without questioning whether the visual tokens themselves carry sufficient spatial concept information.

We argue that the more principled question is whether a vision encoder can produce *compact and structured* visual representations that encode richer spatial concept information per token, without increasing the token budget. This is also well aligned with the goal of visual concept learning, which seeks compact and structured representations of the visual world. Notably, existing VLMs rely almost exclusively on ViT-family encoders, leaving the representation design space largely underexplored.

State space model (SSM) based vision encoders are a promising alternative. Unlike ViTs, which rely on global self-attention over a flattened token sequence, SSM vision models such as VMamba [19] build representations through structured 2D state-space scanning, where each token integrates spatial context from its 2D neighborhood across multiple scan directions. This mechanism embeds spatial relational structure into every token by construction, rather than relying on positional encodings that may be underutilized under standard classification pretraining. To our knowledge, no prior work has conducted a controlled evaluation of SSM encoders as visual backbones in generative VLMs, isolating architectural effects from other confounding factors such as training data, connector design, and pretraining.

In this work, we conduct such a controlled evaluation using a LLaVA-style [18] VLM framework, where we swap only the vision backbone while keeping the rest of the pipeline fixed. Our main findings are: (i) VLMs built on

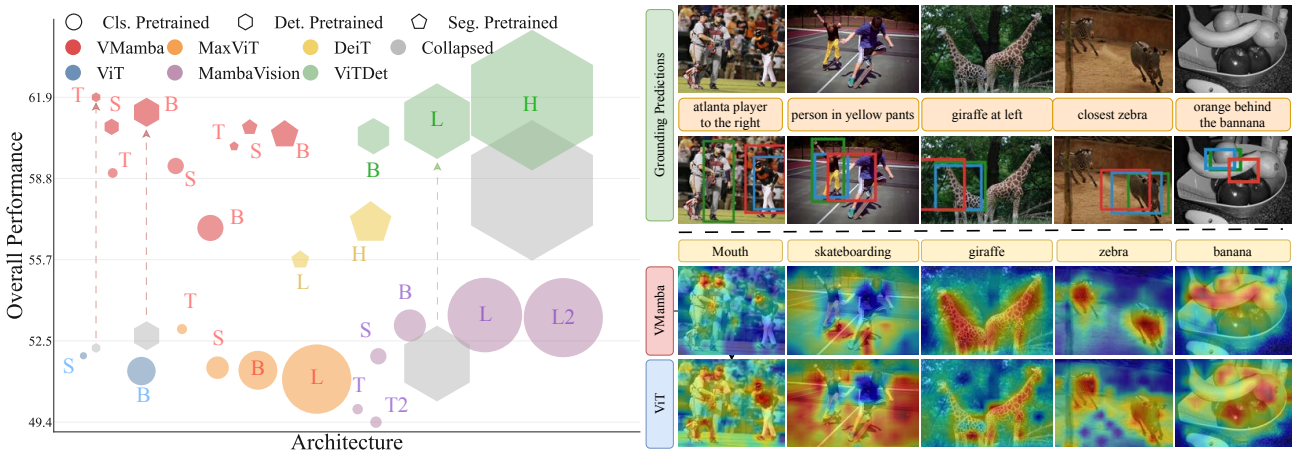

Figure 1. **Left: Overall VLM performance across backbone families.** Each marker represents one backbone configuration plugged into the same VLM pipeline. Colors denote backbone family; marker shapes denote pretraining objective; marker size reflects encoder scale. Gray markers indicate collapsed configurations, with arrows pointing to their stabilized variants. VMamba (red) consistently achieves stronger overall performance than ViT-family alternatives at matched or smaller model scale, despite similar ImageNet accuracy. **Right: Grounding predictions and token-region similarity maps.** Top: predicted bounding boxes from VMamba-T (blue) and ViT-S (red) against ground truth (green). Bottom: similarity maps between visual tokens and text concept tokens computed at an intermediate layer of the language model, showing that VMamba-T tokens produce sharper and more spatially selective concept-region binding than ViT-S tokens. This reveals that the language model can retrieve spatial concept information more effectively from SSM visual representations.

SSM visual representations achieve stronger spatial concept understanding than ViT-family alternatives at matched token budget, with improvements that transfer from grounding benchmarks to general VQA, leading to strong overall performance especially at small model scale (Figure 1, left) (ii) token-region similarity maps computed at intermediate layers of the language model show that SSM visual tokens produce sharper and more spatially selective concept-region binding, providing direct interpretability evidence that the language model can use SSM representations more effectively; (iii) classification pretraining is misaligned with spatial concept retention, and dense pretraining objectives partially recover spatial structure; and (iv) grounding failures in some configurations stem from the vision-language interface rather than the visual representation itself, and can be fixed with simple architecture-agnostic stabilization. In our experiments, we also observe that localization benchmark scores correlate strongly with general VQA scores further supporting this connection. This suggests that improving the spatial concept quality of visual representations is a general strategy for improving VLMs, not only a fix for grounding-specific tasks. Our contributions are:

- We identify compact and structured visual token representations as a key axis for improving spatial concept understanding in VLMs, and argue that architectural inductive bias is an underexplored direction compared to scaling token counts or resolution.
- We provide the first controlled evaluation of SSM-based vision encoders as frozen visual backbones in VLMs, showing that VMamba-based VLMs achieve stronger spatial concept understanding than ViT-family alternatives at matched scale and token budget.
- We provide interpretability evidence through token-region similarity maps at intermediate LLM layers, showing that the language model can leverage SSM visual tokens more effectively for spatial concept reasoning.
- We identify and diagnose two failure modes: classification pretraining eroding spatial concept structure, and vision-language interface bottlenecks causing localization collapse, and propose simple stabilizations for both.

2. Preliminaries

We adopt a LLaVA-style [18] VLM architecture consisting of a frozen vision encoder, a lightweight two-layer MLP connector, and a decoder-only language model (Vicuna-7B). The vision encoder is kept frozen during training; only the connector and language model are updated. This design allows us to swap vision backbones while keeping all other components fixed, enabling a controlled comparison of architectural effects. We evaluate six vision encoder families: VMamba [19], ViT [6], MaxViT [26], MambaVision [8], ViTDet [16], and DeiT [25]. A brief description of each architecture is provided in the supplementary material.

We follow the training recipe of [12], using a one-stage instruction tuning setup on 665K multimodal instruction-tuning examples [18]. Full training details are provided in the supplementary material.

We evaluate on two groups of benchmarks. For general VQA, we use VQA-v2 [7], GQA [10], VizWiz [2], TextVQA [23], POPE [17], and TallyQA [1]. For spatial grounding and localization, we use RefCOCO, RefCOCO+, RefCOCOg [13], and OCID-Ref [28]. We report weighted average scores across each group, as well as an overall weighted average across all benchmarks.

3. SSM Visual Representations Improve VLMs

3.1. Matched Token Budget Comparison

We compare vision encoders from four architecture families under a strictly matched setting: all backbones are pretrained on ImageNet-1K at 224×224 resolution and produce 196 visual tokens. The four families are VMamba [19] (pure SSM), ViT [6] (plain transformer), MaxViT [26] (hybrid convolution and attention), and MambaVision [8] (hybrid SSM and transformer). This setting isolates the effect of backbone architecture from all other factors.

Table 1 reports the weighted average VQA and localization scores for each backbone. VMamba-T/S achieves the strongest performance across both VQA and localization, with a particularly large margin on localization over all other families. Notably, higher ImageNet accuracy does not predict better VLM performance: backbones with the highest classification accuracy consistently underperform VMamba, suggesting that classification-task accuracy is orthogonal to spatial concept quality.

Table 1. **Matched IN1K/224 backbone swaps.** All backbones use 224×224 inputs and L=196 visual tokens. IN1K denotes ImageNet-1K top-1 accuracy. We report weighted average VQA, localization (Loc.), and overall scores. Best results in bold, second best underlined.

Visual Encoder	Size	IN1K	VQA	Loc.	Overall
ViT-S	22M	78.8	57.25	17.82	51.95
ViT-B	87M	81.1	57.17	13.92	51.36
MaxViT-T	31M	83.4	58.75	15.79	52.98
MaxViT-S	69M	84.4	57.42	13.32	51.49
MaxViT-B	119M	84.9	57.60	11.41	51.39
MaxViT-L	212M	84.9	57.30	10.81	51.05
MambaVision-T	32M	82.3	54.38	20.98	49.89
MambaVision-S	50M	83.3	55.61	28.22	51.93
MambaVision-B	98M	84.2	56.53	31.17	53.12
MambaVision-L	228M	85.0	56.94	31.51	53.52
VMamba-T	30M	82.6	<u>62.07</u>	**39.20**	<u>59.00</u>
VMamba-S	50M	83.6	**62.39**	<u>39.17</u>	**59.27**
VMamba-B	80M	83.9	61.38	27.89	56.88

3.2. Interpretability: How LLMs Use Visual Tokens

Figure 1 (right, bottom) shows token-region similarity maps computed between visual tokens and text concept tokens at an intermediate layer of the language model, reflecting both the spatial structure encoded in the visual tokens and the degree to which it is preserved through the language model's intermediate representations.

VMamba tokens produce tight and spatially selective concept-region binding: for example, the concept "Mouth" activates a small, precise facial region, and " zebra" activates the correct animal with minimal spread to surrounding regions. In contrast, ViT tokens produce more diffuse responses, with high similarity spread across multiple regions or objects. This pattern is consistent across all examples in , and aligns with the grounding predictions in Figure 1 (right), where VMamba-based VLMs predict tighter and more accurate bounding boxes than their ViT counterparts. Together, these results suggest that SSM visual tokens carry denser spatial concept structure per token, and that the language model is better able to ground referenced concepts when using SSM representations.

Furthermore, across all evaluated configurations, localization benchmark scores correlate strongly with general VQA scores (Pearson r between 0.65 and 0.80 for VQA-v2, GQA, POPE, and TallyQA), confirming that improvements in spatial concept structure propagate broadly beyond explicit grounding tasks.

4. What Determines Spatial Concept Quality?

4.1. Architectural Inductive Bias

Why does VMamba produce better-structured visual tokens under identical training conditions? VMamba's SS2D mechanism applies state-space aggregation in four scan directions across the 2D token grid, so each token is built by integrating spatial context from its neighbors along rows and columns. Spatial relational structure is therefore embedded into every token by construction. In contrast, ViT's self-attention is permutation-invariant: spatial structure is carried only by positional encodings, which standard classification pretraining does not explicitly reward the model to utilize [11, 27]. As a result, SSM tokens are intrinsically more compact spatial concept carriers per unit of token budget, which explains the sharper concept-region binding observed in Figure 1 (right, bottom).

4.2. Effect of Pretraining Objectives

Classification pretraining optimizes for global category identity and creates a representation bottleneck: models are rewarded for retaining information that distinguishes between object categories, while spatial layout information that is irrelevant for predicting the image label is discarded. This explains why higher ImageNet accuracy does not predict better VLM performance, as seen in Table 1. Adapting backbones with dense pretraining objectives such as detection or segmentation partially recovers spatial concept

fidelity by directly supervising spatial layouts and region-level discrimination. As shown in Table 2, segmentation-adapted VMamba variants consistently outperform DeiT baselines of comparable size on both VQA and localization. For detection-adapted backbones, naive application of these checkpoints causes localization collapse for both ViTDet and VMamba variants (highlighted in blue), which we address in Section 4.3; after stabilization, VMamba achieves the strongest overall performance at a much smaller model scale than ViTDet. The improvement from dense objectives is larger for ViT-family backbones, which lack a built-in spatial prior, than for VMamba, which already preserves spatial structure architecturally, confirming that inductive bias and pretraining objective are complementary levers.

Table 2. **Effect of dense pretraining objectives and interface stabilizations.** We report weighted average VQA, localization (Loc.), and overall scores. Blue rows are collapsed configurations. (f) denotes a stronger 3-layer MLP connector. † denotes square 512×512 input geometry.

Visual Encoder	Size	Obj.	VQA	Loc.	Overall
Detection-adapted (IN1K → COCO)					
ViTDet-B	111M	Det	63.00	43.74	60.42
ViTDet-L	331M	Det	57.64	13.05	51.65
+ (f)	331M	Det	63.55	44.58	61.00
VMamba-S	50M	Det	62.78	47.94	60.78
VMamba-T	30M	Det	58.05	14.86	52.25
+ †	30M	Det	64.22	44.52	61.57
+ (f) + †	30M	Det	64.28	46.75	61.92
VMamba-B	89M	Det	58.57	15.02	52.72
+ †	89M	Det	63.58	45.63	61.17
+ (f) + †	89M	Det	63.58	46.90	61.34
Segmentation-adapted (IN1K → ADE20K)					
DeiT-S	58M	Seg	59.36	31.78	55.65
DeiT-B	134M	Seg	60.69	33.77	57.07
VMamba-T	30M	Seg	62.60	43.47	60.03
VMamba-S	50M	Seg	63.21	44.98	60.76
VMamba-B	89M	Seg	62.87	45.08	60.48

4.3. Vision-Language Interface Bottlenecks

Even when a backbone produces high-quality spatial representations, the language model may still fail to use them if the vision-language interface does not transmit them faithfully. Table 2 shows that naively applying detection-pretrained checkpoints at their original pretraining resolution (1024×1024 for ViTDet, 1333×800 for VMamba-Det) leads to localization collapse in several configurations: ViTDet-L drops to 13.05 and VMamba-T/B drop to around 15, far below their stable counterparts ViTDet-B (43.74) and VMamba-S (47.94). Since the spatial concept information is clearly present in these backbone families, the failure may lie at the interface rather than in the representation itself. We identify two separable bottlenecks: a transmission bottleneck, where connector capacity is insufficient to preserve spatial structure when projecting into the language model embedding space, and a utilization bottleneck, where the language model cannot reliably interpret spatial cues from non-square input geometries. Switching to a square 512×512 input geometry eliminates collapse for VMamba-T and VMamba-B, recovering localization scores from below 15 to above 44. Increasing connector depth provides additional gains and is complementary to the geometry fix, as seen in the VMamba-T(f)† and VMamba-B(f)† rows. Both fixes are architecture-agnostic and add no additional parameters to the vision encoder.

Together, these results show that spatial concept quality is jointly determined by three separable and independently addressable factors: architectural inductive bias, pretraining objective alignment, and vision-language interface fidelity.

5. Related Work

Vision backbone comparisons in VLMs. Several works study the effect of vision encoder choice in VLMs [5, 12, 24]. However, these comparisons often change multiple factors simultaneously, including pretraining data, connector design, and training recipe, making it difficult to isolate the contribution of backbone architecture. We address this by conducting a controlled backbone swap under a fixed training recipe.

SSMs in vision-language models. Prior work applying SSMs to vision-language modeling has focused on the language and fusion side, using Mamba-style layers to replace transformer-based text backbones or multimodal fusion blocks [20, 30], or exploring fully SSM-based contrastive pretraining [22]. None of these works evaluate SSM architectures as frozen vision encoders in generative VLMs under controlled conditions.

6. Conclusion

We have shown that spatial concept understanding in vision-language models is not only a question of how many visual tokens to use, but also of how much spatial concept structure each token carries. SSM-based vision encoders produce visual tokens with richer spatial concept structure per token than ViT-family alternatives at matched scale and token budget, with improvements that transfer from grounding to general VQA. Classification pretraining is misaligned with spatial concept retention, dense objectives partially recover this structure, and grounding failures in some configurations are localized to the vision-language interface and are correctable with simple fixes. We hope this work encourages further exploration of SSM-based vision encoders as a compact and structured alternative to transformer-based backbones.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. 3

[2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 3

[3] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 1

[4] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1

[5] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Llava-more: A comparative study of llms and visual backbones for enhanced visual instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4278–4288, 2025. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 1

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[8] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025. 2, 3, 1

[9] Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. N\" uwa: Mending the spatial integrity torn by vlm token pruning. *arXiv preprint arXiv:2602.02951*, 2026. 1

[10] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 3

[11] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022. 3

[12] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 4

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3

[14] Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Vitoria Lobo, and Mubarak Shah. Weakly supervised grounding for vqa in vision-language transformers. In *European Conference on Computer Vision*, pages 652–670. Springer, 2022. 1

[15] Kevin Li, Sachin Goyal, João D. Semedo, and J Zico Kolter. Inference optimal VLMs need fewer visual tokens and more parameters. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[16] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527, 2022. 2, 1

[17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023. 3

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[19] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 1, 2, 3

[20] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024. 4

[21] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024. 1

[22] Nimrod Shabtay, Itamar Zimerman, Eli Schwartz, and Raja Giryes. Climp: Contrastive language-image mamba pretraining. *arXiv preprint arXiv:2601.06891*, 2026. 4

[23] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[24] 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. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 4

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. arxiv 2020. *arXiv preprint arXiv:2012.12877*, 2(3), 2020. 2, 1

[26] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 2, 3, 1

[27] Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and ZHAO-XIANG ZHANG. Droppos: Pre-training vision transformers by reconstructing dropped positions. *Advances in Neural Information Processing Systems*, 36:46134–46151, 2023. 3

[28] Ke-Jyun Wang, Yun-Hsuan Liu, Hung-Ting Su, Jen-Wei Wang, Yu-Siang Wang, Winston Hsu, and Wen-Chin Chen. Ocid-ref: A 3d robotic dataset with embodied language for clutter scene grounding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5333–5338, 2021. 3

[29] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[30] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10421–10429, 2025. 4

Toward Compact and Structured Visual Representations in VLMs: SSM-Based Vision Encoders as an Alternative to Transformers

Supplementary Material

7. Training Details

Optimization. We use Fully Sharded Data Parallel (FSDP) training on 4× NVIDIA H200 GPUs with BF16 mixed precision and activation checkpointing for the language model. Training is run for 1 epoch with global batch size 128 and per-GPU batch size 16, corresponding to 2 gradient accumulation steps. We use AdamW with learning rate 2×10^{-5}, weight decay 0.1, max gradient norm 1.0, and a linear-warmup cosine-decay schedule with warmup ratio 0.03. These hyperparameters are fixed across all experiments.

Data and preprocessing. We fine-tune on the LLaVA-v1.5 665K instruction-tuning mixture [18], which includes a combination of visual question answering, captioning, and conversation data. Images are processed with letterbox resizing, which preserves the original aspect ratio by scaling and padding to the target resolution. Incomplete batches are retained rather than dropped. For variable-length inputs, text is truncated to the model maximum length and images are padded within each batch as needed.

Implementation. Our implementation is based on the Prismatic-VLMs codebase [12] and uses a fixed batch order across all experiments to ensure reproducibility. One-stage instruction tuning, where the connector is randomly initialized and jointly trained with the language model from the start, is used throughout following [12].

8. Vision Encoder Architectures

VMamba. VMamba [19] is a pure SSM-based vision backbone built around a 2D Selective Scan (SS2D) mechanism. Rather than applying self-attention over a flattened token sequence, SS2D aggregates spatial context in four scan directions across the 2D token grid, embedding spatial relational structure into every token by construction. We use VMamba as our primary SSM backbone given its strong performance on dense vision tasks.

ViT. ViT [6] tokenizes an image into fixed-size patches and applies global self-attention over the full patch sequence. Spatial structure is carried by positional encodings rather than the attention mechanism itself, making it permutation-invariant by design.

MaxViT. MaxViT [26] is a hierarchical hybrid backbone that combines convolutions with multi-axis attention, applying both blocked local attention and dilated global attention at each stage to capture local and global spatial interactions simultaneously.

MambaVision. MambaVision [8] is a hybrid Mamba-Transformer backbone that uses Mamba blocks in early stages for efficient spatial aggregation and retains self-attention in the final layers to capture long-range dependencies.

ViTDet. ViTDet [16] adapts a plain ViT backbone for object detection by showing that a simple feature pyramid built from a single-scale feature map suffices for detection fine-tuning. We use ViTDet checkpoints pretrained on ImageNet-1K and fine-tuned for detection on COCO at 1024×1024 resolution as our ViT-based dense-objective baseline.

DeiT. DeiT [25] is a plain ViT backbone trained with improved data augmentation and distillation. We use DeiT checkpoints adapted with the ViT-Adapter framework [3] for segmentation on ADE20K at 512×512 resolution as our ViT-based segmentation baseline.