CLIMP: Contrastive Language-Image Mamba Pretraining

Nimrod Shabtay[1,2], Itamar Zimerman[1], Eli Schwartz[2], and Raja Giryes[1]

[1] Tel-Aviv University
[2] IBM Research

Abstract. Contrastive Language-Image Pre-training (CLIP) relies on Vision Transformers whose attention mechanism is susceptible to spurious correlations and scales quadratically with resolution. To address these limitations, we present CLIMP, the first fully Mamba-based contrastive vision-language model that replaces both the vision and text encoders with state-space architectures. VMamba's cross-scan mechanism captures spatial inductive biases that reduce reliance on spurious correlations, producing an embedding space with tighter cross-modal alignment and lower hubness - geometric properties that translate to superior retrieval and out-of-distribution robustness, surpassing even CLIP-ViT-B trained on a dataset 167× larger on ImageNet-O. CLIMP naturally supports variable input resolutions without positional encoding interpolation or specialized training, achieving up to 6.6% higher retrieval accuracy at 16× training resolution while using 5× less memory and 1.8× fewer FLOPs. Mamba's autoregressive nature further enables processing of arbitrarily long text, overcoming CLIP's fixed 77-token context limitation for dense captioning retrieval. Our scaling experiments across model sizes and dataset sizes show consistent, unsaturated improvements - indicating that CLIMP's architectural advantages are not limited by training scale. These results demonstrate that Mamba is a compelling alternative to Transformers for vision-language pre-training. The code and models are publicly available at https://github.com/NimrodShabtay/CLIMP

1 Introduction

Contrastive Language-Image Pre-training (CLIP) [40] is a fundamental approach for learning transferable visual representations through natural language supervision. By aligning image and text embeddings in a shared latent space, CLIP enables zero-shot transfer across a range of downstream tasks. However, the Vision Transformer (ViT) [14] backbone commonly employed in CLIP exhibits quadratic computational complexity with respect to sequence length, posing challenges when processing high-resolution images. Moreover, the pairwise token interactions in self-attention can be susceptible to spurious correlations [48, 61], leading to fragile representations under distribution shift.

Mamba [19], a state-space model (SSM), has shown promising results as an alternative to Transformers in sequence modeling. It achieves sub-quadratic

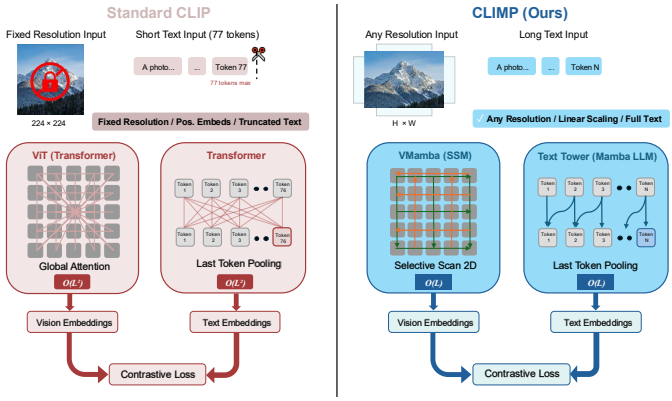

Fig. 1: CLIP vs. CLIMP. By replacing Transformer encoders with Mamba-based models, CLIMP achieves sub-quadratic $O(L)$ complexity instead of quadratic $O(L^2)$. It also removes the fixed resolution and 77-token text limitations of standard CLIP.

complexity in sequence length L during training and constant-time complexity during inference, with these efficiency benefits becoming more pronounced in long-context scenarios. These properties make Mamba an attractive backbone for vision-language tasks requiring long-context processing such as dense captioning.

In the vision domain, adaptations such as Vision Mamba (Vim) [62], Simba [38] and VMamba [34] have applied state space models to image understanding tasks with encouraging results. Beyond computational considerations, which allow high resolution processing, these models introduce spatial inductive biases that differ from the pairwise token interactions characteristic of self-attention. Prior work has shown that such spatial bias benefits vision tasks by enabling more robust [15,36] and sample-efficient learning [16] - properties that ViTs lack. However, the integration of Mamba-based vision encoders into contrastive vision-language frameworks remains relatively unexplored. Prior work [24] has investigated hybrid architectures combining Mamba and Transformer modules within individual towers, but does not explore a fully SSM-based dual-encoder architecture nor provide comprehensive evaluation across retrieval, robustness, resolution flexibility, and dense captioning. We empirically confirm that VMamba's spatial inductive bias transfers to the vision-language setting, enabling improved sample efficiency and robustness (Section 4.5).

In this work, we present Contrastive Language-Image Mamba Pretraining (CLIMP), a fully Mamba-based vision-language model that replaces both the vision and text encoders with state-space architectures within the CLIP framework. We pair VMamba [34] as the vision encoder with Mamba-1 [19] or Mamba-2 [10] language models as text encoders, creating the first end-to-end SSM-based contrastive vision-language model. Our controlled experiments on CC12M [7] isolate architectural differences from data effects, and our scaling experiments across both model and dataset sizes demonstrate that CLIMP's advantages are consistent and unsaturated, indicating they extend beyond our training regime.

Our main contributions are: (1) CLIMP, the first fully Mamba-based CLIP model, integrating VMamba for vision and Mamba LLMs for text. **(2)** Superior retrieval performance driven by tighter cross-modal alignment and reduced hubness in the embedding space. **(3)** Strong out-of-distribution robustness, achieving the highest average accuracy across five ImageNet variants. On ImageNet-O, we surpass CLIP-ViT-B/16 trained on LAION-2B - a dataset 167× larger than ours - suggesting that architectural inductive biases can outweigh data scale for OOD robustness. **(4)** Native variable-resolution support without complex positional encoding schemes or dedicated training, with significantly lower memory and compute overhead. **(5)** Favorable scaling behavior across model sizes (22M–87M parameters) and dataset sizes (1M–12M pairs) with unsaturated performance curves, indicating that CLIMP's advantages are architectural and should persist at larger scales.

2 Related Work

Contrastive Vision-Language Learning. CLIP [40] marked a paradigm shift in vision-language learning by demonstrating that models trained on large-scale image-text pairs can achieve remarkable zero-shot transfer capabilities. It learns a joint embedding space where semantically similar images and texts are mapped close together, enabling open-vocabulary recognition without task-specific fine-tuning. Following CLIP's success, numerous works have sought to improve upon its framework. OpenCLIP [9] provides an open-source reproduction with reproducible scaling laws, training models on datasets such as LAION-400M [45] and LAION-2B [44]. SigLIP [59] replaces the softmax-based contrastive loss with a pairwise sigmoid loss, improving memory efficiency and enabling training with smaller batch sizes. EVA-CLIP [47] enhances the training recipe with stronger vision and text encoders, achieving state-of-the-art zero-shot performance. ALIGN [28] scaled the training data with noisy image-text pairs. Recent work has also explored improving multimodal representations through caption adaptation and knowledge-enhanced supervision [57]. More recent efforts include MetaCLIP [55], which introduces data curation algorithms for balanced training distributions, and SigLIP 2 [51], which adds captioning-based pretraining and self-supervised losses. A recent work has also investigated improving the interpretability of vision-language models through structured semantic representations [58].

Despite these advances, all existing CLIP variants rely on Transformer-based vision encoders, predominantly the Vision Transformer (ViT) [14]. While ViT has proven highly effective, its quadratic complexity with respect to sequence length poses challenges for high-resolution image processing. Scaling ViT to higher resolutions requires substantial computational resources and for a dynamic resolution processing ViT requires a specialized positional encoding schemes such as RoPE [46] or dedicated training scheme [5]. Furthermore, studies have shown that CLIP's robustness to distribution shifts, while impressive on ImageNet variants [17,49], may be overestimated when evaluated on datasets specifically designed to probe spurious correlations [53]. These limitations motivate

our exploration of alternative vision backbones that offer improved efficiency at high resolutions while enhancing robustness.

Recent work explored using LLMs as CLIP text encoders. LLM2CLIP [25] fine-tunes CLIP by replacing its text encoder with a Transformer-based LLM and training a lightweight adaptor, while UniViTAR [39] pairs LLaMA with a ViT-based native-resolution framework. In contrast, CLIMP uses Mamba LLMs as text encoders, enabling a fully SSM-based architecture with consistent sub-quadratic complexity across both modalities - rather than enhancing an existing Transformer-based CLIP or retaining a Transformer vision backbone.

CLIMP is the first work to systematically investigate Mamba vision encoders within the CLIP framework. Replacing ViT with VMamba addresses the computational bottleneck of high-resolution processing and enables native variable-resolution inference that achieves improved zero-shot performance and robustness to distribution shifts.

State Space Models (SSMs) for Vision. SSMs have emerged as a promising alternative to Transformers for sequence modeling. Mamba [19] introduced a selective mechanism making SSM parameters input-dependent, enabling content-aware reasoning with linear complexity. Mamba-2 [10] refined this via the State Space Duality (SSD) framework, linking SSMs and attention while achieving 2-8× speedups.

Adapting Mamba for vision tasks presents unique challenges, as images are inherently 2D and lack the sequential structure of language. Vision Mamba (Vim) [62] addresses this by introducing bidirectional scanning with position embeddings, achieving competitive performance on ImageNet classification. SiMBA [38] addresses the stability issues of Mamba when scaling to large vision networks by introducing Einstein FFT for channel modeling, achieving state-of-the-art SSM performance on ImageNet and multiple time series benchmarks. VMamba [34] proposes the 2D Selective Scan (SS2D) module, which unfolds image patches along four traversal paths to capture global context while maintaining linear complexity. This cross-scan enables each patch to integrate information from all spatial directions, effectively establishing global receptive fields. Subsequent works have extended visual SSMs to various domains, including medical image segmentation [43], video understanding [30], and point cloud processing [31]. Recent studies have also begun exploring the robustness properties of visual SSMs. The works in [15, 36] investigate the robustness of visual SSMs for image classification, finding that Mamba-based architectures exhibit different failure modes compared to ViTs under adversarial perturbations and distribution shifts. These findings suggest that the inductive biases of SSMs may offer complementary advantages to attention-based models in terms of robustness.

While prior work has established the effectiveness of visual SSMs for classification tasks, their potential for vision-language learning remains largely unexplored. CLIMP bridges this gap by integrating VMamba into the CLIP framework, demonstrating that Mamba-based vision encoders can learn effective multimodal representations. Our experiments reveal that the architectural properties of SSMs - particularly their implicit handling of positional information

		Zero-shot Classification		Zero-shot Retrieval	
Vision Tower	Text Tower	Acc@1	Acc@5	IR@5	TR@5
FlexViT-B/16 [5]	LLaMA-3.2 [18]	26.3	57.4	62.4	72.3
NaFlex-B/16 [11]	LLaMA-3.2 [18]	26.1	56.6	61.5	73.3
ViT-B/16 [14]	LLaMA-3.2 [18]	<u>27.3</u>	56.4	62.8	72.9
RoPE-ViT-B/16 [23]	LLaMA-3.2 [18]	<u>27.3</u>	<u>59.0</u>	63.4	72.9
MetaCLIP-ViT-B/16† [55]	LLaMA-3.2 [18]	25.7	56.9	**66.4**	<u>76.2</u>
CLIMP (VMamba-B [34])	Mamba-2 [10]	26.9	**59.0**	65.2	75.3
CLIMP (VMamba-B) [34])	Mamba-1 [19]	**29.6**	58.5	<u>65.5</u>	**77.0**

Table 1: CLIP-Benchmark Results. Average Results for Zero-shot classification and retrieval performance on 31 datasets (28 for classification, 3 for retrieval; full details in the supplementary material). Both CLIMP variants achieve the top retrieval performance among models with comparable pretraining, outperforming all transformer baselines. For classification, the Mamba-1 variant achieves the best Acc@1, while the Mamba-2 variant ties for the best Acc@5. †MetaCLIP-ViT uses a vision encoder pretrained on MetaCLIP-400M [55] (400× larger than IN-1K); despite this advantage, it achieves lower classification accuracy than CLIMP while reaching comparable retrieval.

through scanning patterns - enable native resolution flexibility and contribute to improved robustness on out-of-distribution benchmarks.

3 CLIMP

Mamba offers several advantages over attention models, including (1) improved long-context efficiency and scalability, (2) enhanced robustness, and (3) positional awareness. These long-context capabilities have recently been further theoretically [2] and empirically [3, 4, 35] analyzed proposing various techniques for improved length extrapolation and identifying mechanisms that stabilize recurrent state-space models during long-sequence processing.

Given the above, replacing Transformers in CLIP entirely with Mamba can lead to sub-quadratic memory complexity in both modalities while maintaining improved representation quality. To better understand the suitability of Mamba for CLIP, we discuss its *inductive bias*, which is an important factor for sample-efficiency, out-of-distribution generalization, and representational capacity.

To empirically study this bias, we present positive results in Section 4.5 using synthetic tasks executed in controlled environments. Analytically, prior work has provided evidence that state-space layers exhibit inductive bias toward smoothness and locality compared to Transformers [63]. For Mamba-2, the inductive bias is easier to characterize: the recurrent update uses a per-step transition matrix $\bar{A}$ parameterized as $\bar{A} = \lambda I$, so under standard stability conditions, the influence of a state k steps in the past decays as λ^k, encouraging locality and smoothness. For Mamba-1, the selective mechanism makes $\mathbf{B}$, $\mathbf{C}$, and Δ input-dependent, which precludes a simple closed-form characterization. However, our empirical results in Section 4.5 confirm that both variants exhibit strong spatial inductive bias in practice.

3.1 Preliminaries

State Space Models (SSMs) map an input sequence $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$, according to the following dynamics:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) + \mathbf{D}x(t),$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and $\mathbf{D} \in \mathbb{R}$ are learnable parameters. For discrete sequences, the system is discretized using step size Δ:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t + \mathbf{D}x_t,$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the discretized parameters. Intuitively, h_t acts as a memory summarizing the input history, where $\bar{\mathbf{A}}$ sets the forgetting rate and $\bar{\mathbf{B}}/\mathbf{C}$ control what is written to and read from it. Mamba [19] introduces input-dependent selection by making $\mathbf{B}$, $\mathbf{C}$, and Δ functions of input x_t, enabling content-aware reasoning with linear complexity. Mamba-2 [10] further constrains $\mathbf{A}$ to a scalar times identity matrix, recasting computation as structured matrix multiplications to gain 2-8× speedup.

3.2 Model Architecture

As shown in Figure 1, CLIMP follows CLIP's dual-encoder architecture, mapping images and text into a shared embedding space. Both encoders are Mamba-based, offering consistent sub-quadratic memory scaling across modalities.

Vision Encoder. We use VMamba [34] as the vision encoder. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we divide it into non-overlapping patches of size $P \times P$ and project them into patch embeddings processed through Visual State-Space (VSS) blocks utilizing the SS2D cross-scan mechanism. The hierarchical structure progressively downsamples feature maps through patch merging, with final features mapped to the shared embedding space via a learned projection W_v.

A key advantage is the implicit handling of spatial relationships through scanning patterns, providing a favorable spatial inductive bias (see Section 4.5), allowing CLIMP to process variable input resolutions without positional encoding interpolation, specialized schemes like RoPE [46], or complex training procedures like FlexViT [5] and NaFlex [11].

Text Encoder. We employ two pretrained variants: Mamba-1 [19] (1.4B parameters) and Mamba-2 [10] (1.3B parameters). While both share the selective state-space foundation, Mamba-2 introduces structured state-space duality (SSD) enabling more efficient computation. Unlike Transformers with mature bidirectional encoders [13,33], Mamba models are autoregressive by design. However, this autoregressive formulation is well-suited for contrastive learning: the last-token representation naturally aggregates full-context information, and the absence of a fixed positional encoding enables processing of arbitrarily long text - a key advantage for dense captioning retrieval (Section 4.4). We further analyze the role of the text tower in Table 10.

Vision Tower	Text Tower	IN-V2	IN-R	IN-A	IN-O	Sketch	Avg
FlexViT-B/16	LLaMA-3.2	32.8/62.5	45.4/74.3	13.2/41.2	38.1/68.2	24.6/50.0	30.8/59.2
NaFlex-B/16	LLaMA-3.2	31.4/60.5	44.8/72.5	12.0/36.0	38.5/67.6	23.5/49.1	30.0/57.1
ViT-B/16	LLaMA-3.2	30.0/57.5	42.9/69.7	13.9/37.6	27.7/55.8	21.7/46.1	27.2/53.3
RoPE-ViT-B/16	LLaMA-3.2	34.4/65.4	**47.8**/**76.0**	**16.3**/**46.3**	40.1/69.9	27.4/53.1	33.2/62.1
CLIMP (VMamba-B)	Mamba-2	37.0/67.6	46.6/74.5	15.6/45.5	**49.8**/77.0	27.0/54.2	34.8/63.7
CLIMP (VMamba-B)	Mamba-1	**37.5**/**68.4**	46.2/74.5	15.5/45.3	48.1/**78.8**	**27.5**/**54.4**	**35.2**/**64.3**

Table 2: Out-of-distribution robustness. Evaluation on ImageNet variant datasets, reporting top-1/top-5 accuracy. Both CLIMP variants achieve the top two average scores, with particularly strong gains on ImageNet-O (+9.7%/+8.0% over the best transformer) and ImageNet-V2 (+3.1%/+2.6%).

Given tokenized input $\mathbf{T} = [t_1, t_2, \ldots, t_L]$ with padding mask $\mathbf{m} \in \{0, 1\}^L$, we extract the hidden state at the last non-padding token as the text representation:

$$\mathbf{t}_{\text{raw}} = \mathbf{H}_k, \quad \text{where } k = \max\{i : m_i = 0\} \tag{1}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L]$ denotes hidden states from the last Mamba layer. This last-token pooling is well-suited for Mamba's causal formulation, where each $\mathbf{h}_t$ is computed recurrently, making the last token the only position with full context access. The representation is projected to the shared embedding space via W_t.

4 Results

We evaluate CLIMP across standard CLIP benchmarks (Section 4.1), out-of-distribution robustness (Section 4.2), resolution flexibility (Section 4.3), dense captioning retrieval (Section 4.4), and provide analysis of representational geometry, efficiency, scaling, and spatial inductive bias (Section 4.5).

Experimental Setup. We train all models on CC12M [7] for 10 epochs at 224×224 resolution using AdamW with cosine learning rate schedule (peak LR 5×10^{-5}), batch size 2048, and projection dimension 768. All vision encoders are base-sized (∼86M parameters) initialized from ImageNet-1K [12]. For CLIMP, we use VMamba-B [34] as vision encoder with two text encoder variants: Mamba-1 (1.4B) [19] and Mamba-2 (1.3B) [10]. We compare against three transformer baselines, all using LLaMA-3.2-1B [18] as text encoder: RoPE-ViT [23], FlexViT [5], and NaFlex-ViT [11]. To further strengthen the comparison, we include a MetaCLIP-ViT-B/16 baseline pretrained on MetaCLIP-400M [55] - a dataset 400× larger than ImageNet-1K.

Our evaluation follows the controlled-comparison paradigm adopted in recent vision-language research [32]: all models are trained on the same dataset with identical protocols, isolating architectural differences from confounding factors such as dataset scale or curation. Our scaling experiments (Section 4.5) confirm that CLIMP's advantages are consistent across data and model sizes with unsaturated performance curves, indicating the scalability of our results.

		Acc@1/Acc@5			IR@5/TR@5		
Vision	**Text**	224	320	384	224	320	384
FlexViT	LLaMA	26.3/57.3	26.1/57.3	25.9/56.9	62.4/72.3	63.5/74.3	63.6/75.0
NaFlex	LLaMA	26.1/56.6	25.5/56.6	25.1/56.1	61.4/73.2	63.0/74.1	63.4/74.6
RoPE-ViT	LLaMA	27.3/**59.0**	26.4/**58.4**	24.9/**57.1**	63.3/72.9	64.1/74.3	63.3/73.2
VMamba	Mamba-2	26.9/**59.0**	26.1/58.2	24.9/56.9	65.2/75.2	**66.2**/76.2	**65.2**/74.6
VMamba	Mamba-1	**29.6**/58.5	**28.4**/57.5	**27.3**/56.3	**65.5**/**77.0**	66.0/**76.9**	65.0/**75.2**

Table 3: Resolution scaling on CLIP-Benchmark. All models support arbitrary resolution inference. Both CLIMP variants demonstrate strong performance across all resolutions, achieving the best results on Acc@1 and retrieval metrics (IR@5/TR@5).

4.1 CLIP-Benchmarks Results

We evaluate on CLIP-Benchmark [8] (31 datasets for zero-shot classification and retrieval). Table 1 presents our main findings. Both CLIMP variants achieve top retrieval performance: Mamba-1 leads with 65.5% image and 77.0% text recall, outperforming the best transformer baseline (RoPE-ViT) by +2.1% and +4.1% respectively. For classification, CLIMP-Mamba-1 achieves the best Acc@1 (29.6%, +2.3% over the best transformer), while CLIMP-Mamba-2 ties for best Acc@5 (59.0%). Notably, MetaCLIP-ViT - despite vision pretraining on 400× more data - achieves lower classification accuracy (25.7%) and comparable retrieval, suggesting that CLIMP's advantages stem from architectural properties rather than data scale. We further stress the architectural gains by comparing CLIMP to convolution based variants (i.e CLIP based ResNets [20]), We evaluate on Open-CLIP [27] 38 benchmarks, CLIMP outperformed all ResNet based models more than 4 points on average on both zero-shot classification and retrieval, outperforming the strongest ResNet variant. Full details are in the supplementary materials.

Between our two variants, Mamba-1 consistently outperforms Mamba-2 on most metrics. While Mamba-2 achieves 2–8× computational speedups through key simplifications - restricting the transition matrix from a structured diagonal (HiPPO-initialized) to a scalar identity, and adopting multi-head weight sharing - these reduce per-token expressivity. Mamba-1's per-channel selective scan, with fully input-dependent parameters (Δ, $\mathbf{B}$, $\mathbf{C}$), provides finer-grained filtering that has more discriminative representations for contrastive alignment.

4.2 OOD Robustness

Having established CLIMP's retrieval advantages, we next examine whether its representations also improve robustness under distribution shift. We evaluate on five ImageNet variants: ImageNet-V2 [42], ImageNet-R [21], ImageNet-A [22], ImageNet-O [22], and ImageNet-Sketch [52].

Table 2 shows that CLIMP variants achieve the top two average robustness scores, with Mamba-1 leading at 35.2% top-1 and 64.3% top-5, outperforming the best transformer baseline (RoPE-ViT) by +2.0% and +2.2% respectively.

		NoCaps						Crossmodal-3600					
		224		512		896		224		512		896	
Vision	**Text**	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
RoPE-ViT	LLaMA	67.6	78.9	65.1	**75.9**	38.6	34.9	64.4	68.2	62.3	**64.1**	36.8	26.9
NaFlex	LLaMA	60.4	69.1	59.7	65.0	51.6	50.8	56.7	59.3	56.3	55.7	48.2	42.0
FlexViT	LLaMA	56.0	62.9	53.4	57.8	42.2	41.2	52.6	50.5	50.9	47.6	40.4	34.1
CLIMP(VMamba)	Mamba-2	70.2	81.3	66.4	72.4	**57.1**	**53.8**	**65.2**	**69.2**	**62.8**	61.9	54.7	**46.0**
CLIMP(VMamba)	Mamba-1	**70.5**	**81.9**	**66.9**	72.7	57.1	53.5	65.1	69.1	62.6	60.8	**55.9**	45.4

Table 4: Retrieval Recall@5 (IR@5/TR@5) on NoCaps [1] and Crossmodal-3600 [50] at different resolutions for both image to text (I→T) and text to image (T→I). Both CLIMP variants demonstrate strong performance across all resolutions, with the performance gap over transformer baselines widening substantially at 896×896.

The most striking result is on ImageNet-O. Both CLIMP variants dramatically outperform all transformer baselines. Mamba-2 achieves 49.8% and Mamba-1 48.1% top-1 accuracy - surpassing the best transformer by +9.7% and +8.0%. Remarkably, both variants also surpass CLIP-ViT-B-16 [40] (42.3%)[3] trained on LAION-2B [44], a dataset 167× larger than CC12M. This provides direct evidence that architectural inductive biases can outweigh data scale for OOD robustness, consistent with our geometric analysis (Section 4.5) showing that CLIMP's reduced hubness leads to less reliance on spurious feature correlations.

Both CLIMP variants also lead on ImageNet-V2 (+3.1%/+2.6%) and Sketch. On ImageNet-R and ImageNet-A, RoPE-ViT performs best, though CLIMP remains competitive within 1–2%, consistent with prior findings that SSMs and transformers exhibit complementary failure modes [15].

4.3 High-Resolution

We next evaluate resolution flexibility - a practical requirement for applications that process images at their native resolution. Unlike transformers that require positional embedding interpolation (RoPE-ViT) or specialized training (FlexViT, NaFlex), CLIMP generalizes to higher resolutions without architectural modifications or additional training. We train all models at 224×224 and evaluate at up to 896×896 with no fine-tuning.

CLIP-Benchmark Evaluation. We evaluate on CLIP-Benchmark at resolutions up to 384×384, matching the native resolution range of its constituent datasets. Table 3 shows that both CLIMP variants achieve the top retrieval performance across all resolutions, with Mamba-1 achieving the best Acc@1 throughout. This advantage persists at 320×320 and 384×384 without any resolution-specific adaptation.

Scaling to Higher Resolutions. To rigorously evaluate resolution scaling beyond standard benchmarks, we test on NoCaps [1] (4.5K images, avg. 810×960) and Crossmodal-3600 [50] (3.6K images, avg. 640×520) - benchmarks whose native resolutions substantially exceed 224×224. While performance decreases

[3] Taken from: OpenCLIP Results (March 2026)

		Image Retrieval			Text Retrieval		
Vision Tower	**Text Tower**	R@1	R@5	R@10	R@1	R@5	R@10
FlexViT	LLaMA	46.4	75.0	84.5	55.7	79.0	87.8
NaFlex	LLaMA	50.9	80.2	88.9	66.1	86.8	93.8
ViT	LLaMA	53.9	82.0	90.1	64.8	85.8	92.0
RoPE-ViT	LLaMA	60.7	85.8	93.3	77.6	93.6	97.0
CLIMP (VMamba)	Mamba-2	63.3	88.2	94.6	75.7	92.6	96.2
CLIMP (VMamba)	Mamba-1	**67.0**	**89.4**	**94.7**	**81.3**	**95.9**	**98.4**

Table 5: Zero-shot retrieval on rephrased Flickr8k-test with dense captions. To make the captions dense, the original captions were rephrased using an LLM (avg. 134 tokens vs. original 14), with 98.32% of them exceeding the 77-token limit. Our CLIMP models consistently outperform transformer-based baselines, demonstrating effective retrieval with extended textual descriptions beyond CLIP's 77-token limit.

		Image Retrieval			Text Retrieval		
Vision Tower	**Text Tower**	R@1	R@5	R@10	R@1	R@5	R@10
RoPE-ViT	LLaMA	18.0	39.5	50.5	10.5	27.9	38.1
FlexViT	LLaMA	20.5	42.8	53.8	15.6	36.3	47.6
NaFlex	LLaMA	29.1	55.3	67.5	**26.1**	**53.4**	**65.1**
CLIMP (VMamba)	Mamba-2	33.4	62.6	74.0	24.6	51.7	63.7
CLIMP (VMamba)	Mamba-1	**37.3**	**67.1**	**77.8**	23.8	50.4	62.7

Table 6: Zero-shot retrieval on DOCCI [37]. DOCCI is a dense captioning dataset with long descriptions (avg. 142 tokens, 94.4% exceeding 77 tokens). Our CLIMP models achieve the best image retrieval across all metrics. On text retrieval, CLIMP remains competitive but NaFlex leads, likely benefiting from its sequence packing design.

at higher resolutions for all models - as expected when evaluating at 2–4× the training resolution - CLIMP degrades far more gracefully. Table 4 shows retrieval performance across resolutions. Both CLIMP variants outperform all baselines at every resolution, with the gap widening substantially at higher resolutions. At 896×896, CLIMP achieves +18–19% advantage over RoPE-ViT on both image and text retrieval. Notably, both variants surpass FlexViT and NaFlex despite these models being specifically designed for variable resolutions.

4.4 Dense Captioning Retrieval

Complementing our high-resolution evaluation, we evaluate robustness to out-of-distribution text lengths. CLIMP overcomes the standard CLIP's fixed 77-token context window through Mamba's autoregressive text encoder, which naturally handles arbitrarily long sequences. Note that in our evaluation, no method is restricted to 77 tokens at evaluation - all transformer baselines use LLaMA-3.2-1B (8K native context) and Mamba LLMs support arbitrary length, so long captions are processed in full by every model. We evaluate on: (1) Flickr8k-test

Model	Data	Train Loss	Test Acc.
VMamba	Regular	**1.028**	**69.33**
ViT	Regular	1.121	67.50
VMamba	Shuffled	1.612	46.44
ViT	Shuffled	**1.328**	**59.72**

Table 7: Spatial inductive bias analysis. We train lightweight 3-layer VMamba (0.33M params) and ViT (0.35M params) on CIFAR-10 with regular and shuffled patch orders. VMamba achieves lower training loss on regular images but degrades significantly under shuffling, while ViT shows the opposite pattern—performing relatively better on distorted data. This confirms that SSM-based encoders rely on sequential spatial structure as an inductive bias.

captions rephrased using LLaMA-3.3-70B [18] (average 134 tokens, 98.3% exceeding 77 tokens), with images at 224×224, and (2) DOCCI [37] with naturally verbose descriptions (average 142 tokens, 94.4% exceeding 77 tokens), with images at 896×896. As shown in Tables 5 and 6, CLIMP-Mamba-1 achieves the best image retrieval across both benchmarks, with up to 6.3% improvement on Flickr8k-Rephrased and 8.2% on DOCCI over the best transformer baseline. On DOCCI text retrieval, NaFlex achieves the highest scores, likely benefiting from its sequence packing design; CLIMP remains competitive on this metric while substantially leading on image retrieval.

4.5 Analysis

Spatial Inductive Bias To understand the source of VMamba's advantages, we investigate its spatial inductive bias. Prior work has shown that locality bias benefits vision tasks by enabling more sample-efficient learning [16], a property ViTs possess only weakly through their positional encoding (PE). We validate this using lightweight 3-layer VMamba (0.33M params) and ViT (0.35M params) models trained on CIFAR-10 [29] with regular versus spatially distorted data with shuffled patch orders. VMamba achieves lower training loss on regular images, while ViT performs better on the distorted variant, confirming that SSM-based encoders encode sequential spatial structure as a much stronger inductive bias. Full results appear in Table 7.

Qualitative Results CLIMP produces more interpretable image-text alignment maps. As shown in Figure 2, CLIMP correctly localizes the wooden deck and fence structure described in the caption, while RoPE-ViT and FlexViT exhibit diffuse attention scattered across irrelevant areas.

Representational Geometry Analysis To understand CLIMP's retrieval advantages, we analyze embedding geometry on NoCaps [1]. Following [54], we measure *alignment* (matched pairs should be close), *uniformity* (embeddings

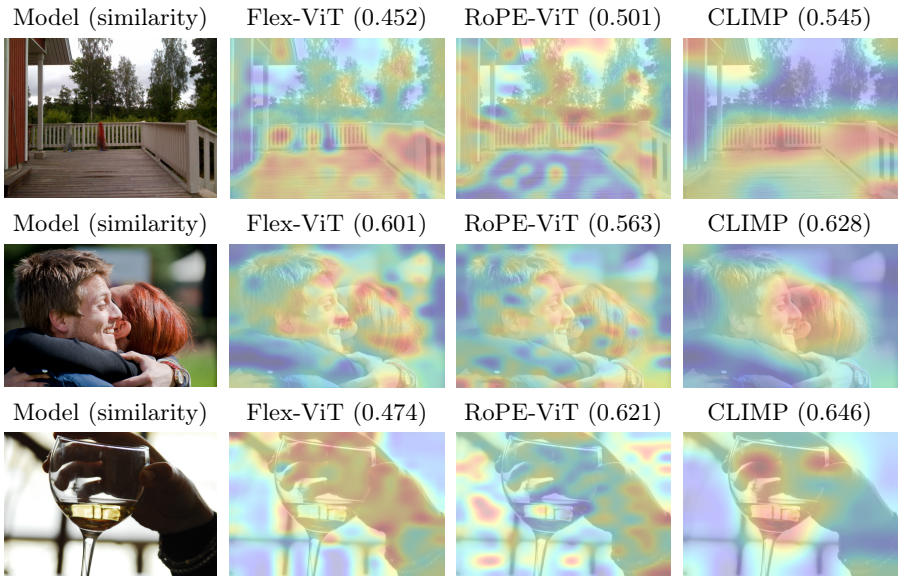

Fig. 2: Visualization of Image-Text similarity for the captions: "A large porch with a wooden fence and no roof." (top), "A woman with red hair hugging a man." (middle) and "A hand holding a glass that has some wine in it." (bottom). CLIMP (similarity: 0.545) produces spatially coherent attention focused on the porch and fence, while RoPE-ViT (0.501) and FlexViT (0.452) show scattered, less interpretable patterns. Warmer colors indicate higher similarity.

spread evenly on the hypersphere), and *hubness* [41] - the tendency for certain embeddings to dominate nearest-neighbor lists, degrading retrieval quality [60].

Table 8 shows that CLIMP achieves the best alignment (0.982) and lowest text hubness (1.13), directly explaining its retrieval gains: better alignment improves matching accuracy, while reduced text hubness benefits text-to-image retrieval. On image-side metrics, NaFlex achieves the best image uniformity and hubness, suggesting complementary strengths across architectures. The lower text hubness also implies less reliance on generic, spuriously correlated features, which may contribute to CLIMP's improved OOD robustness (Section 4.2).

CLIMP vs. Transformer-CLIP: Qualitative Insights A fair comparison requires articulating not only where CLIMP wins, but where transformers retain an edge. Using the CLIP-Benchmark suite, we organize both sides by the architectures' internal mechanisms and prior analyses of visual SSMs. **CLIMP outperforms transformers in:** (1) *High resolution.* The largest and most consistent advantage: the gap widens with input resolution, and by 896×896 CLIMP exceeds RoPE-ViT on retrieval, driven by Mamba's native variable-resolution support. (2) *Sparse, distributed perturbations.* CLIMP's scanning-based receptive field dilutes distributed spurious cues, while attention latches onto them locally; consistent with the ImageNet-O gain, see 2. **Transformers may be favor-**

			Uniformity ↓		Hubness ↓	
Vision Tower	**Text Tower**	**Alignment ↓**	Image	Text	Image	Text
FlexViT	LLaMA	1.111	-2.87	-3.50	1.19	1.23
NaFlex	LLaMA	1.039	**-3.30**	-3.50	**0.93**	1.24
RoPE-ViT	LLaMA	1.052	-3.16	-3.52	1.04	1.25
CLIMP (VMamba)	Mamba-1	**0.982**	-3.18	-3.54	1.01	1.15
CLIMP (VMamba)	Mamba-2	1.000	-3.16	**-3.54**	1.01	**1.13**

Table 8: Embedding geometry analysis on NoCaps. We analyze alignment, uniformity, and hubness. CLIMP achieves better alignment and lower text hubness than transformer baselines, explaining its superior retrieval performance. ↓ = lower is better for all metrics.

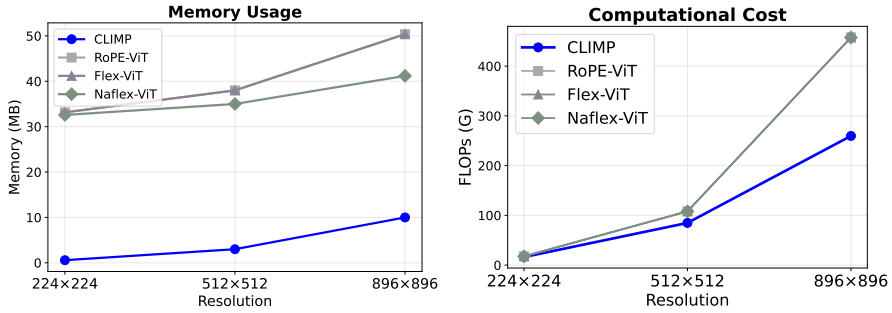

Fig. 3: Efficiency analysis. CLIMP achieves superior memory and computational efficiency across all resolutions. (Left) Memory overhead is 4–57× lower. (Right) FLOPs scale linearly, yielding up to 1.8× reduction—a gap that widens with resolution.

able in: (1) *Low-resolution inputs.* Few patches mean fewer scanning steps for context aggregation, while ViT's global attention is comparatively unaffected. (2) *Strong visual-style shift.* VMamba's spatial features may transfer less well than ViT's abstract attention features, consistent with RoPE-ViT's slight lead on ImageNet-R. Visual examples appear in the supplementary materials.

Memory and Computational Efficiency CLIMP offers significant efficiency advantages at high resolutions. As shown in Figure 3, CLIMP requires 5× less resolution-specific memory (10.0 vs 50.4 MB) and 1.8× fewer FLOPs (259.7 vs 457.9 GFLOPs) than ViT variants at 896×896, with the gap widening at higher resolutions due to Mamba's linear complexity.

Scaling We investigate how CLIMP scales with model size and training data, directly addressing whether its advantages persist at larger scales.
Model Size. Table 9 shows ImageNet-1K zero-shot classification across model scales (22M–87M parameters). CLIMP consistently outperforms ViT-based alternatives at every scale, with the advantage maintained from the smallest (22M:

Params	CLIMP (Mamba2)	FlexViT (LLaMA)	RoPE-ViT (LLaMA)
22-30M	38.8	32.5	33.2
50M	42.8	–	–
87M	43.5	38.3	40.7

Table 9: Scaling behavior on ImageNet-1K zero-shot classification. CLIMP consistently outperforms ViT-based alternatives across all scales, with performance improving steadily as model size increases.

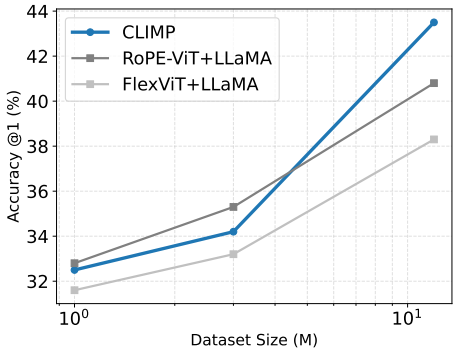

Fig. 4: CLIMP scaling laws trained on CC and evaluated on ImageNet-1K Acc@1. Performance improves consistently as training data scales from 1M to 12M samples, with no sign of saturation, indicating that our findings should scale further.

+5.6% over FlexViT) to the largest (87M: +2.8% over RoPE-ViT) configuration.

Dataset Size. Figure 4 examines scaling with training data on Conceptual Captions. CLIMP exhibits a steep scaling curve that has not saturated at 12M samples, suggesting continued benefits with larger datasets. The consistent performance gap over transformer baselines across all data sizes (1M, 3M, 12M) indicates that CLIMP's advantages stem from architectural properties rather than overfitting to a particular data regime. This is further supported by the MetaCLIP-ViT comparison (Table 1): despite vision pretraining on 400× more data, it achieves lower classification accuracy than CLIMP, providing direct evidence that our architectural advantages extend beyond the training scale.

4.6 Ablations

We ablate the text encoder contribution by replacing the LLM backbone while keeping the vision encoder fixed. As shown in Table 10, the results reveal an architectural synergy: while RoPE-ViT shows minimal sensitivity to text encoder choice (±1% across metrics), VMamba consistently benefits from Mamba-based text encoders, achieving +3.5% Acc@1 and +3.9% TR@5 over LLaMA. This suggests that matched SSM architectures in both modalities learn more compatible representations for cross-modal alignment. We additionally evaluate Hydra [26],

Vision Tower	Text Tower	Acc@1/Acc@5	IR@5/TR@5
RoPE-ViT [23]	LLaMA [18]	27.27/58.95	63.35/72.93
RoPE-ViT [23]	Qwen2-1.5B [56]	27.30/58.80	60.40/70.90
RoPE-ViT [23]	Mamba-2 [10]	27.42/58.69	62.59/72.99
RoPE-ViT [23]	Mamba-1 [19]	28.10/58.54	63.17/74.28
RoPE-ViT [23]	RoBERTa-L [33]	27.47/58.64	64.82/74.72
RoPE-ViT [23]	BERT-L [13]	27.80/58.84	63.93/74.15
VMamba [34]	LLaMA [18]	26.13/57.30	63.21/73.16
VMamba [34]	Hydra [26]	28.63/56.98	61.24/71.35
VMamba [34]	Mamba-2 [10]	26.94/**59.04**	65.17/75.25
VMamba [34]	Mamba-1 [19]	**29.59**/58.53	**65.54**/**77.03**

Table 10: Architectural synergy. We compare language model backbones across vision encoders. While RoPE-ViT shows minimal sensitivity to text encoder choice, VMamba consistently benefits from Mamba-based text encoders, suggesting that matched SSMs learn more compatible cross-modal representations. Hydra [26], a bidirectional SSM, underperforms the autoregressive Mamba LLMs.

a bidirectional SSM variant, which underperforms the autoregressive Mamba LLMs - confirming that mature pretrained language models provide stronger text representations than current bidirectional SSM alternatives.

5 Conclusions

We introduced CLIMP, the first contrastive vision-language model built entirely on Mamba. By replacing ViT with VMamba and pairing it with Mamba LLMs, CLIMP achieves linear complexity in both modalities while producing superior representation quality and robustness. Our experiments reveal: (1) superior retrieval performance, driven by tighter cross-modal alignment and reduced hubness; (2) improved OOD robustness, notably surpassing CLIP-ViT-B/16 trained on LAION-2B on ImageNet-O; (3) memory and FLOPs reductions at high resolutions with native variable-resolution support; (4) dense captioning retrieval beyond CLIP's token limit; (5) spatial inductive bias contributing to sample-efficient learning; and (6) favorable scaling behavior across model and dataset sizes, with unsaturated performance curves indicating that these advantages are architectural and should persist at larger scales.

These findings establish SSMs as a viable alternative to transformers for vision-language pre-training, overcoming limitations of CLIP implementations for high-resolution, memory-constrained, and long-context applications. Future directions include scaling to larger models and datasets, extending to generative vision-language tasks, and developing hybrid architectures [6].

6 Acknowledgments

This work was partially supported by Tel Aviv University Center for AI and Data Science (TAD) and Blavatnik Interdisciplinary Cyber Research Center (ICRC).

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
2. Bar, N., Seleznova, M., Alexander, Y., Kutyniok, G., Giryes, R.: Revisiting glorot initialization for long-range linear recurrences. In: Conference on Neural Information Processing Systems (NeurIPS) (2025)
3. Ben-Kish, A., Zimerman, I., Abu-Hussein, S., Cohen, N., Globerson, A., Wolf, L., Giryes, R.: Decimamba: Exploring the length extrapolation potential of mamba. In: International Conference on Learning Representations. pp. 101148–101170 (2025)
4. Ben-Kish, A., Zimerman, I., Mirza, M.J., Wolf, L., Glass, J.R., Karlinsky, L., Giryes, R.: Overflow prevention enhances long-context recurrent LLMs. In: Conference on Language Modeling (2025)
5. Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., Pavetic, F.: Flexivit: One model for all patch sizes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14496–14506 (2023)
6. Blakeman, A., Grattafiori, A., Basant, A., Gupta, A., Khattar, A., Renduchintala, A., Vavre, A., Shukla, A., Bercovich, A., Ficek, A., et al.: Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. arXiv preprint arXiv:2512.20848 (2025)
7. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
8. Cherti, M., Beaumont, R.: Clip benchmark (Nov 2022). `https://doi.org/10.5281/zenodo.15403103`, `https://doi.org/10.5281/zenodo.15403103`
9. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
10. Dao, T., Gu, A.: Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In: Proceedings of the 41st International Conference on Machine Learning. pp. 10041–10071 (2024)
11. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., et al.: Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. Advances in Neural Information Processing Systems **36**, 2252–2274 (2023)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.:

An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
15. Du, C., Li, Y., Xu, C.: Understanding robustness of visual state space models for image classification. arXiv preprint arXiv:2403.10935 (2024)
16. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International conference on machine learning. pp. 2286–2296. PMLR (2021)
17. Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L.: Data determines distributional robustness in contrastive language image pre-training (clip). In: International conference on machine learning. pp. 6216–6234. PMLR (2022)
18. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
19. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
21. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
22. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.X.: Natural adversarial examples. 2021 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15257–15266 (2019)
23. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024)
24. Huang, W., Shen, Y., Yang, Y.: Clip-mamba: Clip pretrained mamba models with ood and hessian evaluation. arXiv preprint arXiv:2404.19394 (2024)
25. Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., Dai, Q., Wang, C., Dai, X., Chen, D., et al.: Llm2clip: Powerful language model unlocks richer visual representation. arXiv preprint arXiv:2411.04997 (2024)
26. Hwang, S., Lahoti, A., Puduppully, R., Dao, T., Gu, A.: Hydra: Bidirectional state space models through generalized matrix mixers. Advances in Neural Information Processing Systems **37**, 110876–110908 (2024)
27. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
29. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
30. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European conference on computer vision. pp. 237–255. Springer (2024)
31. Liu, J., Yu, R., Wang, Y., Zheng, Y., Deng, T., Ye, W., Wang, H.: Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. arXiv preprint arXiv:2403.06467 (2024)

32. Liu, Y., Zhang, Y., Ghosh, D., Schmidt, L., Yeung-Levy, S.: Data or language supervision: What makes clip better than dino? arXiv preprint arXiv:2510.11835 (2025)
33. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
34. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
35. Lu, P., Huang, J., Zeng, Q., Wang, X., Chen, B., Langlais, P., Cui, Y.: Mamba modulation: On the length generalization of mamba models. In: Conference on Neural Information Processing Systems (NeurIPS) (2025)
36. Malik, H.S., Shamshad, F., Naseer, M., Nandakumar, K., Khan, F.S., Khan, S.: Towards evaluating the robustness of visual state space models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3544–3553 (2025)
37. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., Wang, S., Baldridge, J.: DOCCI: Descriptions of Connected and Contrasting Images. In: ECCV (2024)
38. Patro, B.N., Agneeswaran, V.S.: Simba: Simplified mamba-based architecture for vision and multivariate time series. arXiv preprint arXiv:2403.15360 (2024)
39. Qiao, L., Gan, Y., Wang, B., Qin, J., Xu, S., Yang, S., Ma, L.: Univitar: Unified vision transformer with native resolution. ArXiv (2025)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of machine learning research **11**(sept), 2487–2531 (2010)
42. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
43. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
44. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
45. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
46. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
47. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
48. Tamayo-Rousseau, C., Zhao, Y., Zhang, Y., Balestriero, R.: Your attention matters: to improve model robustness to noise and spurious correlations. arXiv preprint arXiv:2507.20453 (2025)

49. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems **33**, 18583–18599 (2020)
50. Thapliyal, A.V., Tuset, J.P., Chen, X., Soricut, R.: Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 715–729 (2022)
51. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
52. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
53. Wang, Q., Lin, Y., Chen, Y., Schmidt, L., Han, B., Zhang, T.: A sober look at the robustness of clips to spurious features. Advances in Neural Information Processing Systems **37**, 122484–122523 (2024)
54. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
55. Xu, H., Xie, S., Tan, X., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. In: International Conference on Learning Representations (2024)
56. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z.: Qwen2 technical report (2024), `https://arxiv.org/abs/2407.10671`
57. Yanuka, M., Ben-Kish, A., Bitton, Y., Szpektor, I., Giryes, R.: Bridging the visual gap: Fine-tuning multimodal models with knowledge-adapted captions. In: Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL). pp. 10497–10518 (2025)
58. Yellinek, N., Karlinsky, L., Giryes, R.: 3vl: Using trees to improve vision-language models' interpretability. IEEE Transactions on Image Processing **34**, 495–509 (2025)
59. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
60. Zhang, T., Suya, F., Jha, R., Zhang, C., Shmatikov, V.: Adversarial hubness in multi-modal retrieval. arXiv preprint arXiv:2412.14113 (2024)
61. Zhou, Y., Zhu, Z.: Fighting spurious correlations in text classification via a causal learning perspective. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 4264–4274 (2025)
62. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning. pp. 62429–62442. PMLR (2024)
63. Zimerman, I., Wolf, L.: Viewing transformers through the lens of long convolutions layers. In: Forty-first International Conference on Machine Learning (2024)