SAGIVTECH IMVC 2026
15th Israel Machine Vision Conference (IMVC) 2026
Monday | April 27, 2026 | Pavilion 10, EXPO Tel Aviv

CLIMP: Contrastive Language-Image Mamba Pretraining

Nimrod Shabtay, Tel Aviv University

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 167x 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 16x training resolution while using 5x less memory and 1.8x 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.