Let ViT Speak: Generative Language-Image Pre-training

Yan Fang[1,2,*], Mengcheng Lan[2,3,*], Zilong Huang[2,†], Weixian Lei[2], Yunqing Zhao[2], Yujie Zhong[2], Yingchen Yu[2], Qi She[2], Yao Zhao[1], and Yunchao Wei[1,†]

[1]Beijing Jiaotong University, China [2]ByteDance, China
[3]Nanyang Technological University, Singapore
{yanfang, yunchao.wei}@bjtu.edu.cn
lanm0002@e.ntu.edu.sg zilong.huang2020@gmail.com
Home: https://github.com/YanFangCS/GenLIP

Abstract. In this paper, we present **Gen**erative **L**anguage-**I**mage **P**re-training (GenLIP), a simplified generative pretraining approach for Vision Transformers (ViTs) tailored to multimodal large language models (MLLMs). To better align ViTs with the autoregressive nature of LLMs, GenLIP trains a ViT to predict language tokens directly from visual tokens using a standard language modeling objective—without contrastive batch construction or an additional text decoder. GenLIP offers three key advantages: (1) **Simplicity**: a single transformer jointly models visual and linguistic tokens; (2) **Scalability**: it scales efficiently with both data and model size; (3) **Performance**: it achieves competitive or superior results across diverse multimodal benchmarks. With only 1/5 of the seen samples, GenLIP matches or surpasses strong baselines such as SigLIP2. With continued pretraining on multi-resolution images at native aspect ratios, GenLIP further excels at detail-sensitive tasks such as OCR, chart understanding, and visual question answering, making it a strong foundation for vision encoders in MLLMs.

1 Introduction

Multimodal Large Language Models (MLLMs) have emerged as a transformative paradigm in artificial intelligence, demonstrating remarkable capabilities in understanding and reasoning across vision and language modalities [6, 12, 43, 60, 83]. The prevailing architecture of MLLMs comprises three core components: a vision encoder for processing visual information [13, 20, 55, 78], a connector for bridging modalities, and a large language model (LLM) as the reasoning engine [1, 5, 64, 66]. Among these components, the vision encoder serves as the *perceptual foundation*, responsible for extracting meaningful visual representations that can be effectively consumed by the downstream LLM. The quality and design of this vision encoder fundamentally determine the upper bound of an MLLM's visual understanding capabilities. Currently, large-scale Vision-Language Pre-training (VLP) on billions of image-text corpora have become the dominant approach for developing strong vision encoders.

*Equal Contribution. † Corresponding authors.

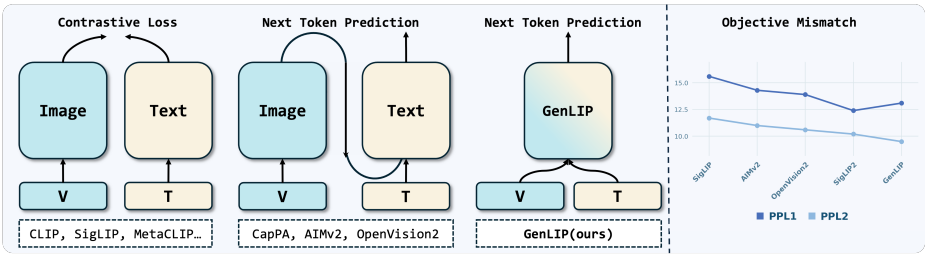

Fig. 1: Overview of GenLIP and compatibility analysis. Left: Compared with prior vision-language pretraining methods that rely on dual-encoder structures or additional text modules, GenLIP adopts a simplified architecture that directly trains visual tokens with next-token prediction. We use "V" and "T" to denote visual and textual inputs, respectively. **Right:** we connect different vision encoders to an LLM through a projector and measure perplexity when tuning only the projector (PPL1) or tuning both the projector and LLM (PPL2). Lower perplexity indicates better compatibility with the LLM's generative objective.

Contrastive learning based VLP methods, exemplified by CLIP [55] and SigLIP [78], are among the most widely adopted vision encoders in MLLMs [9, 62, 63]. These methods typically employ a dual-encoder architecture that encodes each modality separately and align them using a contrastive objective. However, contrastive pretraining may not be fully aligned with the generative training paradigm of MLLMs: while contrastive learning encourages discriminative image-text alignment, MLLMs are optimized through next-token prediction. Consistent with this mismatch, the diagnostic evaluation in Figure 1 shows that vision features from generative objectives tend to yield lower perplexity after being attached to an LLM, suggesting better compatibility with the LLM's generation.

Another stream of works focus on generative pretraining, such as CapPa [68], AIMv2 [22], and OpenVision2 [45]. These methods typically cascade the vision encoder with a text decoder, and train the cascaded model using an autoregressive language modeling objective. In this setup, the vision encoder is updated indirectly via gradients flowing through the text decoder. Related hybrid designs, such CoCa [77] and SigLIP2 [67], further introduce an additional text encoder to combine contrastive and generative objectives. While these approaches narrow the gap, the redundant architecture design and indirect optimization can limit efficiency for training a scalable vision encoder in MLLMs.

To unleash the full potential of generative vision-language pretraining, we advocate for a simple design philosophy: remove unnecessary modules and train the vision backbone as directly as possible. In this work, we propose a simplified framework for generative vision-language pretraining: **Gen**erative **L**anguage-**I**mage **P**retraining (GenLIP), a simple yet scalable framework that departs from the complex designs of prior VLP methods. Instead of introducing novel architectural components, our core insight is elegantly simple: *let the Vision Transformer (ViT) speak directly*–requiring no contrastive batch construction and no additional text module.

Instead of indirectly optimizing vision transformers through auxiliary text modules, GenLIP directly pretrains the ViT to generate language tokens that describe the visual content using only a standard autoregressive language modeling objective. This simple generative paradigm naturally aligns the vision encoder with the way LLMs operate, mitigating the objective mismatch in contrastive approaches while simplifying designs and improving efficiency for generative frameworks. GenLIP's design philosophy offers several compelling advantages: **(1) Alignment:** GenLIP optimizes the vision encoder with an autoregressive token prediction objective, matching the generative training paradigm in downstream MLLMs; **(2) Simplicity:** GenLIP uses a minimalist form of a single vision backbone and an autoregressive objective, without contrastive losses or additional text modules; **(3) Scalability:** GenLIP scales effectively with data and model size, yielding strong empirical performance. Across extensive experiments, GenLIP matches or outperforms strong baselines pretrained on much larger corpora while using only 8B pretraining samples, and its second-stage native-aspect-ratio adaptation further improves downstream performance.

In summary, GenLIP provides a direct and efficient formulation of generative vision-language pretraining. Our results suggest that a simpler and better-aligned pretraining paradigm can serve as a strong foundation for future MLLMs. We believe these findings chart a more direct, efficient, and scalable course for developing powerful vision-language models.

2 Related Works

The convergence of computer vision and natural language processing has been driven by large-scale vision-language pretraining, which aims to learn robust, generalizable multimodal representations from massive image-text corpora. Typical VLP methods can be grouped into three categories based on architectural design and training objectives: dual-encoder contrastive pretraining, encoder-decoder generative pretraining, and simplified single-transformer pretraining.

Dual-Encoder Contrastive Pretraining. A broad line of research has investigated Contrastive Language-Image Pretraining. CLIP-style architectures [13, 30,55,74,78] are fundamentally based on a dual-encoder (two-tower) design, which learns to align image and text representations within a shared embedding space using an InfoNCE or similar contrastive objective. Subsequent works improve alignment by leveraging high-quality image-text pairs [14, 21, 24, 32, 39, 76, 81] or dense region-level captions [38, 41, 79] for fine-grained representation learning. While effective for discriminative tasks such as classification and retrieval, contrastive pretraining primarily focuses on global alignment and does not facilitate deep cross-modal interaction.

Encoder-Decoder Generative Pretraining. To enable richer cross-modal reasoning, recent works [3, 22, 45, 69, 71] adopt generative pretraining, typically cascading a vision encoder with a text decoder. For example, Aimv2 [22] couples a vision encoder with a multimodal decoder that autoregressively generates raw image patches and text tokens, whereas CapPa [68], GIT [69] and OpenVision 2 [45] stack a text decoder on top of the image encoder and pre-train the model

using only a captioning loss. Most recently, some studies [36, 37, 39, 44, 67, 77] form hybrid pretraining schemes that combine a contrastive dual-encoder for image-text alignment with a generative decoder for captioning.

Discussion. Despite their success, existing methods often rely on multiple towers or multiple optimization objectives, which increases model complexity and limits efficiency. Moreover, alignment is often performed at later stages rather than within the image encoder itself, which can constrain early cross-modal interactions. In contrast, we propose a simple generative vision-language pretraining framework with a simplified architecture and training objective–a single transformer and a single language modeling objective.

Single-Transformer Pretraining. Recently, some works also explored vision-language pretraining under a simplified single-Transformer architecture with different objectives. Among them, SuperClass [27] proposes vision transformer pretraining with a single Transformer tower using token-level classification targets. VL-BEiT [8] and OneR [29] aim to unify vision-language representation learning within a single-tower Transformer, but still rely on multiple objectives. Beyond vision transformer pretraining, several recent efforts [11, 17–19, 33, 61] aim to build native MLLMs with a single transformer and a single objective.

Discussion. In particular, GenLIP is architecturally close to SAIL [33], as both use a single transformer with a language modeling objective. However, SAIL focuses on building a native MLLM with a simplified architecture based on pretrained LLMs, whereas GenLIP is designed to pretrain a scalable vision encoder from scratch to better serve modular MLLMs [7, 12, 34]. This distinct goal also leads to different design choices.

3 Methods

This section details GenLIP, our simple implementation for generative vision-language pretraining. We first introduce the core designs of our approach, including the model architecture, data representation, and training objective, all designed for our simplified generative vision-language pretraining. We then provide pretraining details, including pretraining datasets and training schedule.

3.1 GenLIP Framework

Instead of introducing novel architectural components, GenLIP is built upon a simple unified modeling paradigm for vision encoder pretraining. Specifically, we build GenLIP with a simple transformer architecture in the spirit of *letting the ViT speak directly*, analogous to how LLMs generate text. We keep the design simple, introducing only minimal but necessary modifications for improving visual representations.

Data Format. All pretraining data for GenLIP is structured as image-text pairs, denoted as $\{(I_i, T_i)\}_{i=1}^{N}$, where each image I_i is associated with its caption T_i. Each image I_i is partitioned into a sequence of non-overlapping patches $\{v_0, v_1, ..., v_M\}$ using a convolutional patch embedding layer, as in standard ViT models. The corresponding text T_i is tokenized into a sequence of subword tokens

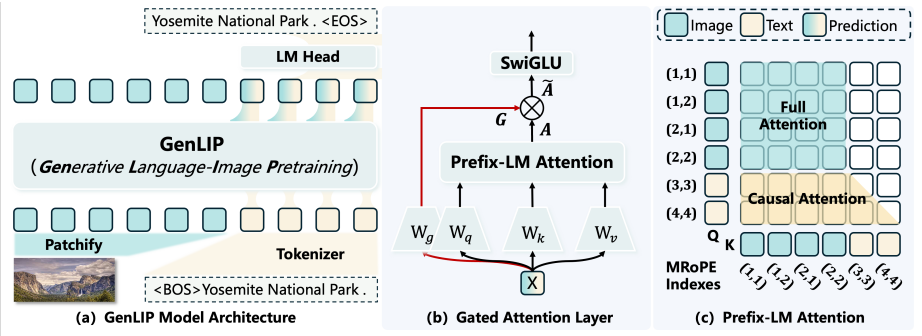

Fig. 2: GenLIP framework. **(a)** A single Transformer processes visual-prefix sequences. **(b)** Gated attention modulates attention outputs. **(c)** Prefix-LM attention gives bidirectional attention to image tokens and causal attention to text tokens.

$\{t_0, t_1, ..., t_L\}$ using an off-the-shelf text tokenizer (Qwen3 [75]). The resulting image patch embeddings and text token embeddings are concatenated into a single sequence, with the image embeddings preceding the text embeddings. The final input sequence S for a given pair (I_i, T_i) is:

$$S = [v_0, \ldots, v_M, t_0, \ldots, t_L]. \tag{1}$$

Architecture. The architecture of GenLIP is designed for simplicity and effectiveness, centered around a unified Transformer encoder that processes a concatenated sequence of image and text tokens. As illustrated in Figure 2, the model consists of three components: modality-specific embedding layers, a unified Transformer with a prefix-LM attention implementation, and a language modeling (LM) head for token prediction.

To enable effective cross-modal interactions and unified modeling of the interleaved multimodal sequence, we make two small but crucial modifications to a standard Transformer. (i) To better encode the position information in an interleaved multimodal sequence, we use multimodal rotary position encoding (MRoPE) [70] and discard the absolute position embeddings for image patches. (ii) We replace the basic full attention with prefix-LM attention [56] in all transformer blocks, where image tokens attend bidirectionally and text tokens attend causally. Based on the above two modifications, we directly apply the GenLIP architecture to process the unified multimodal sequence, without additional modality-specific designs in the network architecture.

Objective. GenLIP adopts a single standard autoregressive language modeling objective, applied exclusively to the textual part of the sequence. The model is trained to predict the next text token conditioned on the preceding image and text tokens, thereby directly modeling the conditional probability distribution $P(T|I)$. The objective is to minimize the negative log-likelihood of the text sequence:

$$\mathcal{L}_{\text{LM}} = -\sum_{k=0}^{L} \log P(t_k | \{v_j\}_{j=0}^{M}, \{t_i\}_{i=0}^{k-1}; \theta) \tag{2}$$

where θ denotes the model parameters to be optimized, and $P(t_k|\{v_j\}_{j=0}^{M}, \{t_i\}_{i=0}^{k-1})$ is the predictive probability of the k-th text token conditioned on all preceding visual and textual tokens.

Using GenLIP as a Vision Encoder. When employing GenLIP as a visual encoder, we extract vision features from the output of the LN layer following the last Transformer block and feed them into a 2-layer MLP projector to align them with the LLM's input space. In this process, the language modules of GenLIP (the tokenizer and LM head) are discarded because no text inputs are used, while all other components are retained. The Prefix-LM attention mechanism reduces to standard full attention when GenLIP is used as a vision encoder.

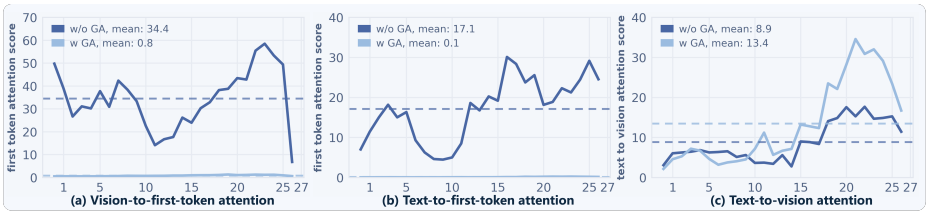

Fig. 3: Layer-wise attention allocation in controlled So/16 models. We analyze the attention allocation of different source tokens across 27 layers. The X-axis represents the layer index. We report the attention mass from (a) vision tokens to the first sequence token, (b) generated text tokens to the first sequence token, and (c) generated text tokens to vision tokens. Dashed lines denote the layer-averaged attention scores. GA effectively reduces the first-token attention sink in both vision and text streams, while strengthening text-to-vision attention in deeper layers.

3.2 Gated Attention

While the above unified architecture is effective for generative vision-language pretraining, we observe a notable side effect: attention becomes overly concentrated on the first token of the input sequence, a phenomenon known as the *attention sink*. This issue is particularly pronounced in our mixed-modality setting, as shown in Figure 3. Under full attention, certain visual tokens can freely aggregate global information from all patches, effectively becoming image-level summaries. Since text tokens only access visual information through causal attention over this shared visual prefix, the model learns a shortcut: compressing visual information into a few sink tokens for efficient language prediction, at the cost of degrading spatial diversity in visual representations. Consistent with findings in [53], this leads to (i) obvious loss spikes during pretraining, and (ii) attention distributions where the first token absorbs most of the attention mass, reducing the effective utilization of visual tokens. As a result, the pretrained ViT exhibits substantially degraded discriminative performance, such as in ImageNet linear probing, and shows unstable scaling behavior–both undesirable for our target usage as a vision encoder for MLLMs.

Inspired by [53], we introduce a gated attention mechanism to regulate information flow in the mixed-modality modeling space. Given input hidden

states $X \in \mathbb{R}^{n \times d}$ for a Transformer block, we compute a standard attention output $A = \mathrm{Attn}(X)$ and apply a per-head input-dependent gate:

$$G = \sigma(XW_g + b_g), \quad \widetilde{A} = G \odot A, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, W_g and b_g are learnable parameters, and $\odot$ denotes element-wise multiplication. The gated attention output $\widetilde{A}$ is then used in the standard residual pathway. By modulating attention outputs on a per-token basis, the gate prevents text tokens from collapsing their attention onto a small subset of visual tokens and encourages the model to leverage spatially distributed visual features. In practice, gated attention alleviates loss spikes, accelerates convergence, and stabilizes scaling behavior.

3.3 Pretraining Details

Our pretraining comprises two stages with different datasets and resolutions, progressing from fixed low-resolution inputs to diverse resolutions and native aspect ratios. This setup allows the model to learn foundational visual and linguistic representations while keeping the overall computational cost manageable.

Table 1: Overview of GenLIP configurations and pretraining setup. **Left:** model configurations. **Right:** two-stage pretraining details.

Model configurations.

Model	Params	Layers	Dims	Heads	FFN-w
GenLIP-L	0.3B	24	1024	16	2752
GenLIP-So	0.4B	27	1152	16	3072
GenLIP-g	1.1B	40	1536	24	4096

Two-stage pretraining details.

Stage	Dataset	Size	Resolution	Patches	Samples
S1	Recap-DataComp-1B	1.0B	224	196	8.0B
S2	Blip3o-Long-Cap	27M	AnyRes	16-1024	39M
	Infinity-MM-Stage1	10M			
	CapRL	2M			

Fixed Resolution Pretraining. We first pretrain GenLIP on Recap-DataComp-1B [40], a large-scale dataset of 1 billion unique image-text samples collected from the web. During this stage, we use the fixed 224×224 resolution images to reduce computational cost while learning strong foundational visual representations. We train GenLIP for a total of 8 billion samples in this stage, corresponding to 8 epochs over the dataset.

Diverse Resolution Adaptation. We further fine-tune fixed-resolution pretrained GenLIP on public open-source caption datasets, namely the caption subset of Infinity-MM (stage1) [26], BLIP3o-Long-Caption [10] and CapRL [73], containing 39 million image-text samples with long captions and higher-resolution images. Different from higher resolution adaptation in previous works [52, 67], we process images in their native aspect ratios, and resize them to keep the number of vision tokens within $[16, 1024]$. In this adaptation stage, we train GenLIP for only 1 epoch over the datasets. This stage helps the model adapt to variable-resolution inputs and learn finer-grained visual representations from dense text descriptions, which is important for downstream tasks that require detailed visual understanding and precise image-text grounding.

Regularization. We apply two regularization techniques during GenLIP pretraining to effectively train deeper networks: layer scale and drop path. These techniques mainly stabilize training and prevent divergence when training deeper models, although we find that limited impact on the final performance.

3.4 Discussion

Rather than introducing novel architectural components, GenLIP pursues the simplest possible paradigm for vision encoder pretraining, enabling seamless integration into MLLMs. Here, we summarize the key differences between GenLIP and prior works.

Differences from previous generative works. GenLIP differs from previous generative vision-language pretraining works [8, 22, 45, 68] in several key aspects: (i) Compared with VL-BEIT [8] and AIMv2 [22], GenLIP only learns from a single standard autoregressive language modeling objective, without masked image modeling or pixel reconstruction objectives; (ii) Compared with CapPa [68], AIMv2 [22], and OpenVision2 [45], GenLIP discards the additional text decoder and uses a simplified modeling paradigm with a single unified transformer.

Differences from previous single Transformer pretraining works. GenLIP also differs from previous single Transformer pretraining works [18, 33]: (i) GenLIP focuses on pretraining a scalable vision encoder for modular MLLMs, rather than native MLLMs; (ii) GenLIP is pretrained from scratch on caption datasets, while SAIL [33] and NEO [18] are trained by leveraging pretrained LLMs and large-scale instruction-tuning data; (iii) GenLIP improves attention implementation with a gated mechanism to make it better fit visual modeling as a vision encoder.

4 Experiments

To comprehensively evaluate the visual features learned by GenLIP, we first describe the experimental setup and benchmark suite, then report frozen representation and standard LLaVA-NeXT evaluations on multimodal understanding tasks. We next analyze scalability with respect to both pretraining data and model size. Finally, we conduct controlled comparisons and ablations on the pretraining method, SAIL-style architecture and initialization, gated attention, and native-aspect-ratio adaptation.

4.1 Setup

Baselines. We compare our method, GenLIP, against a suite of representative vision-language pre-training models. This includes contrastive methods such as CLIP [55], SigLIP [78], and SigLIP2 [67], as well as generative approaches like AIMv2 [22] and OpenVision2 [45]. For a fair comparison, all vision encoders are configured to produce the same number of visual tokens (patches). We use strong publicly available model variants for our baselines, such as ViT-L/14 for CLIP and AIMv2, and ViT-So/16 for SigLIP2. These methods are pretrained on substantially larger corpora (12.0B–40.0B image-text pairs) than GenLIP.

Experimental Setup. Following Cambrian [65], we mainly adopt frozen visual representation evaluation, where the vision encoder is kept frozen and the language model is fine-tuned on downstream tasks. This protocol directly measures the quality of visual features learned by different VLP methods without the confounding effect of further fine-tuning the vision encoder. Based on the LLaVA-NeXT framework [42], we replace the original vision encoder with one pretrained by GenLIP or each baseline method, and then fine-tune the language model on an instruction tuning dataset. To better unleash the potential of the vision encoders, we replace the original 780K instruction-tuning set with the comprehensive LLaVA OneVision [34] dataset, which contains more than 3 million supervised fine-tuning (SFT) samples. We consider two LLM backbones of different sizes in our implementation, Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct [54], in place of the original LLM in LLaVA-NeXT.

Evaluation Benchmarks. To provide a comprehensive evaluation, we assess our method and all baselines across a diverse set of multimodal understanding benchmarks. These benchmarks are grouped into three categories to probe distinct capabilities: document understanding and optical character recognition (Doc&OCR), general visual understanding (General VQA), and image captioning (Caption). All evaluations are conducted using the open-source LMMS-Eval toolkit [80].

Document and OCR. This category evaluates the model's ability to recognize and interpret text within images, a critical skill for document analysis and scene text understanding. Following mainstream MLLMs [7, 34], we focus on a wide range of classic benchmarks, including ChartQA [49], OCRBench [46], InfoVQA [50], AI2D [31], TextVQA [59], DocVQA [51] and SEED-Bench-2-Plus [35].

General Visual Understanding. This group of tasks assesses the model's broader capabilities in comprehending and reasoning about visual content. We employ four widely-used benchmarks, including MME [23], GQA [28], VQAv2 [25], and ScienceQA [47] for general VQA.

Image Captioning. To measure the model's ability to generate descriptive text from images, we evaluate on NoCaps [2], COCO [48], and TextCaps [57] for evaluation. Performance is reported using the CIDEr metric.

4.2 Main Results

We provide all frozen visual representation evaluation results on multimodal understanding benchmarks in Table 2 and Table 3. Besides, we also provide results under standard LLaVA-NeXT evaluation setting in Table 4.

Frozen feature analysis. As presented in Table 2, GenLIP demonstrates strong performance across three model scales. Despite using fewer pretraining pairs, GenLIP achieves consistent gains over all baselines, including the 40B-pair pretrained SigLIP2. Under the Qwen2.5-1.5B setting, GenLIP improves the overall average (ALL AVG) over SigLIP2 by 2.8, 2.0, and 3.7 points at the L/16, So/16, and g/16 scales, respectively. The gains are especially pronounced on

Table 2: Frozen visual representation evaluation under LLaVA-NeXT-Qwen2.5-1.5B. Benchmarks are grouped into Doc&OCR, General VQA, and Caption tasks.

Model	Arch	Data	Doc&OCR							General VQA				Caption			ALL AVG
			ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	
CLIP [55]	L/14	12.8B	24.8	23.7	38.9	43.9	64.5	30.2	47.8	46.1	39.8	75.3	1218	55.5	72.5	117.9	53.1
AIMv2 [22]	L/14	12.0B	26.3	25.2	37.7	47.2	64.2	29.3	47.3	**48.1**	**43.9**	76.2	1157	80.1	73.6	122.4	55.7
OVision2 [45]	L/16	12.8B	30.7	45.6	43.3	49.2	65.6	28.1	47.8	44.0	42.7	75.5	1230	**84.3**	**76.3**	127.4	58.7
SigLIP [78]	L/16	40.0B	30.2	41.0	47.3	36.0	66.4	27.8	47.9	41.3	41.7	76.7	1203	84.0	76.1	120.7	56.9
SigLIP2 [67]	L/16	40.0B	33.4	45.7	45.1	50.3	**66.7**	28.2	45.7	43.1	42.6	**76.9**	1165	82.9	74.6	127.8	58.7
GenLIP	L/16	8.0B	**41.2**	**51.1**	**51.1**	**53.6**	66.6	**30.7**	**51.1**	44.4	41.5	76.1	**1258**	82.6	76.0	**131.4**	**61.5**
SigLIP2 [67]	So/16	40.0B	35.2	47.2	46.4	53.3	67.0	28.0	50.3	**46.5**	43.5	**77.1**	**1220**	84.3	77.1	**131.5**	60.6
GenLIP	So/16	8.0B	**40.8**	**51.5**	**51.9**	**55.2**	**67.2**	**31.9**	**52.3**	**46.5**	**44.0**	76.0	1215	**87.5**	**81.5**	129.5	**62.6**
SigLIP2 [67]	g/16	40.0B	35.3	47.3	47.6	54.7	66.7	29.7	49.6	**50.1**	45.2	76.2	**1284**	84.4	76.2	134.5	61.5
GenLIP	g/16	8.0B	**45.0**	**55.6**	**57.0**	**59.0**	**68.9**	**33.9**	**53.3**	49.1	**45.5**	**77.5**	1256	**88.3**	**82.0**	**135.4**	**65.2**

Table 3: Frozen visual representation evaluation under LLaVA-NeXT-Qwen2.5-7B. Settings follow Table 2 except for the LLM size.

Model	Arch	Data	Doc&OCR							General VQA				Caption			**ALL AVG**
			ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	
CLIP [55]	L/14	12.8B	36.6	29.6	48.4	52.6	76.3	**39.0**	55.1	49.4	39.6	85.2	1316	63.1	54.4	127.9	58.8
AIMv2 [22]	L/14	12.0B	36.8	30.9	46.6	54.5	76.9	37.5	55.1	44.0	37.9	85.2	1240	66.5	55.9	130.5	58.6
OVision2 [45]	L/16	12.8B	42.5	49.9	49.5	58.8	78.4	33.8	53.8	**60.0**	47.2	85.9	**1325**	79.4	69.6	133.8	64.9
SigLIP [78]	L/16	40.0B	41.7	45.7	50.5	56.0	79.3	34.8	55.8	57.8	46.2	**86.7**	1275	**81.5**	**72.0**	131.1	64.5
GenLIP	L/16	8.0B	**52.7**	**59.2**	**61.7**	**62.9**	**80.4**	38.8	**59.0**	56.4	**51.3**	85.4	1320	81.1	71.3	**139.4**	**69.0**
SigLIP2 [67]	So/16	40.0B	46.6	55.6	56.3	63.5	**81.3**	37.2	56.4	**64.5**	52.2	**87.1**	1422	**84.1**	**76.4**	139.3	69.4
GenLIP	So/16	8.0B	**55.3**	**63.5**	**66.3**	**65.7**	81.0	**41.4**	**60.8**	60.5	**52.4**	86.4	**1424**	83.1	74.8	**142.1**	**71.8**
SigLIP2 [67]	g/16	40.0B	47.2	55.6	56.3	63.5	**81.0**	36.4	56.4	62.7	49.3	**87.7**	1422	82.0	72.3	142.7	68.9
GenLIP	g/16	8.0B	**57.1**	**65.9**	**69.0**	**66.8**	**81.0**	**43.6**	**61.1**	**64.4**	**54.5**	87.0	**1483**	**85.0**	**75.5**	**144.8**	**73.6**

Doc&OCR benchmarks, which demand fine-grained document understanding and text-centric visual reasoning. Averaging over the seven Doc&OCR tasks in Table 2, GenLIP achieves 49.3, 50.1, and 53.2 at L/16, So/16, and g/16, outperforming SigLIP2 by 4.3, 3.3, and 5.9 points, respectively.

This advantage remains under a larger LLM. As shown in Table 3, scaling the LLM to Qwen2.5-7B yields consistent trends with the Qwen2.5-1.5B setting. Under this setting, GenLIP outperforms SigLIP2 by 2.4 and 4.7 points on average score at the So/16 and g/16 scales, respectively. Similar to the Qwen2.5-1.5B setting, GenLIP consistently performs best on Doc&OCR benchmarks, highlighting its strong visual-text alignment.

Across both frozen settings, GenLIP not only surpasses contrastive VLMs such as CLIP [55] and SigLIP [78], but also outperforms prior encoder-decoder generative VLMs, including AIMv2 [22] and OpenVision2 [45]. These generative baselines use an additional text decoder for language modeling, and OpenVision2 is further pretrained on the stronger Recap-DataComp-1B v2 corpus with a longer

Table 4: Multimodal understanding results under standard LLaVA-NeXT settings with the same data and LLM.

Patches	Model	Arch	Data	Doc&OCR						General VQA						ALL AVG
				ChartQA	DocVQA	TextVQA	OCR-B	LiveVQA	AI2D	MMBench	MME-C	MME-P	POPE	RWQA	MMStar	
576	CLIP [55]	L/14	12.8B	75.2	66.5	62.5	52.5	47.4	73.2	74.6	48.0	75.6	88.8	63.7	49.0	64.8
	MLCD [4]	L/14	12.0B	76.5	67.8	61.7	53.1	48.4	77.0	76.5	54.1	79.9	88.7	61.1	51.0	66.3
	AIMv2 [22]	L/14	12.8B	77.2	72.7	65.9	57.2	47.3	75.4	**78.6**	48.3	75.0	88.4	62.2	50.2	66.5
	RICE-ViT [72]	L/14	13.0B	79.2	72.3	65.9	57.5	**48.9**	77.9	76.6	**54.6**	**80.7**	88.5	63.1	51.8	68.1
	GenLIP	So/16	8.0B	**79.3**	**75.2**	**68.5**	**59.7**	48.4	**78.6**	77.7	48.6	78.2	**89.2**	**65.9**	**53.1**	**68.5**
729	SigLIP [78]	So/14	40.0B	76.7	69.3	64.7	55.4	48.4	76.2	77.0	46.1	79.9	88.8	63.7	47.3	66.1
	SigLIPv2 [67]	So/14	40.0B	79.1	70.2	66.2	58.7	48.6	77.0	77.1	46.6	**80.4**	89.3	63.4	52.8	67.5
	RICE-ViT [72]	L/14	13.0B	82.6	75.1	66.2	58.8	49.5	76.5	77.6	54.1	79.0	89.1	62.9	51.2	68.6
	GenLIP	So/16	8.0B	**83.0**	**76.9**	**69.6**	**64.7**	**50.4**	**79.1**	**78.1**	**54.5**	80.1	**89.4**	**65.1**	**53.2**	**70.3**

training schedule. Overall, the results suggest that GenLIP's simple architecture and objective can yield stronger visual representations with improved data efficiency.

We also observe that GenLIP scales favorably with model size, while SigLIP2 shows comparatively smaller gains when scaling up. These results support two hypotheses: (i) simplifying both the architecture and the objective can enable more efficient scaling; and (ii) larger model capacity helps GenLIP learn both broad visual knowledge and fine-grained alignment.

Standard LLaVA-NeXT evaluation. We further evaluate GenLIP under the standard LLaVA-NeXT setting following prior work [72], where the vision encoder is unfrozen and fine-tuned jointly with the language model during instruction tuning. As shown in Table 4, GenLIP performs strongly under two fixed patch budgets and achieves competitive overall results across both Doc&OCR and General VQA tasks.

Taken together, both the frozen and standard evaluations indicate that GenLIP provides strong and consistent performance across diverse multimodal understanding tasks, including Doc&OCR, General VQA, and captioning. In particular, GenLIP consistently excels on Doc&OCR tasks, which demand fine-grained visual recognition and precise visual-text alignment.

Overall, these results indicate that GenLIP, a simple yet effective generative vision-language pretraining method, can learn rich and versatile visual representations for multimodal understanding with high data efficiency. Compared with more complex alternatives (e.g., SigLIP2 with larger pretraining corpora and more elaborate training recipes), GenLIP exhibits highly competitive and often achieves better downstream performance. This suggests that simple generative vision-language pretraining is a promising direction for learning strong, scalable visual representations for MLLMs.

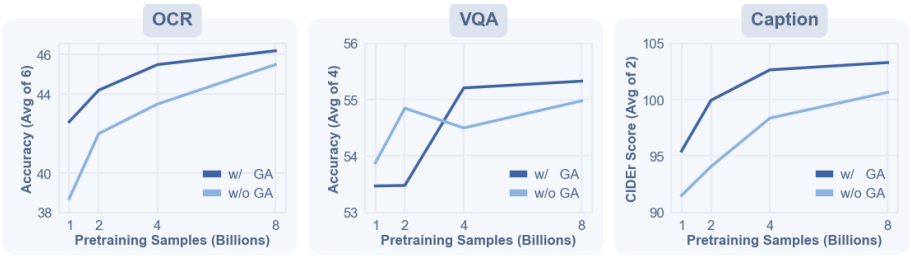

Fig. 4: Data Scaling Behavior. Average Doc&OCR, VQA, and Caption scores as first-stage pretraining scales from 1.0B to 8.0B samples.

Table 5: Frozen visual representation evaluation of GenLIP across model scales after fixed 224×224 pretraining on the same 8.0B samples from the first stage.

Model	Arch	Doc&OCR							General VQA				Caption			ALL AVG
		ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	
GenLIP	L/16	34.3	28.9	44.5	43.0	64.5	28.5	49.1	44.0	41.9	75.2	1136	77.3	71.0	114.1	55.2
GenLIP	So/16	**37.6**	39.2	49.8	50.5	65.3	**29.7**	**51.3**	**45.4**	**43.8**	75.2	1157	80.9	73.6	125.7	58.9
GenLIP	g/16	34.6	**42.5**	**53.7**	**53.1**	**65.5**	29.6	51.1	45.3	43.5	**75.9**	**1164**	**82.0**	**74.0**	**132.0**	**62.0**

4.3 Scalability Analysis

To investigate the detailed scaling pattern of GenLIP, we discuss both data and model scalability of GenLIP, which are two key factors for VLP pretraining.

Data Scaling. We first study data scaling in Fig. 4, where we pretrain GenLIP (with or without gated attention) on the 1B pretraining caption dataset with different numbers of training samples, ranging from 1.0B to 8.0B. As the data scale increases from 1.0B to 8.0B, GenLIP shows sustained improvements on multimodal understanding benchmarks. We observe steeper gains when scaling from 1.0B to 4.0B, while the improvement curve becomes flatter when further scaling to 8.0B. In particular, the average performance on VQA and caption tasks shows only minor improvements when scaling from 4.0B to 8.0B. Based on this trend, we use 8.0B samples as the default data scale for GenLIP pretraining.

Model Scaling. We also investigate how GenLIP performance changes with model size by pretraining GenLIP at the L/16, So/16, and g/16 scales. Besides the final results after diverse resolution adaptation shown in Table 2, we additionally provide results for models pretrained only with fixed low resolution on the first stage caption dataset in Table 5. Across both pretraining stages, GenLIP shows consistent performance gains with increasing model size. An important observation is that GenLIP-L/16 lags behind GenLIP-So/16 and GenLIP-g/16 only with fixed low-resolution pretraining, while the gap between g/16 and So/16 is relatively small. This suggests that an appropriate model size is important for GenLIP to learn strong visual representations for downstream performance.

Table 6: Ablation between different pretraining methods.

Model	Arch	Data	OCR							General VQA				Caption			
			ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	ALL AVG
SigLIP	So/16	2.0B	26.1	36.2	38.6	44.3	64.2	25.8	46.0	42.7	39.8	75.1	1132	76.2	70.6	119.2	54.4
OVision2	So/16	2.0B	27.8	**43.2**	41.2	44.7	64.1	26.8	46.3	44.2	40.3	74.8	**1158**	76.3	71.0	**123.3**	55.9
GenLIP	So/16	2.0B	**35.0**	36.9	**46.0**	**47.1**	**64.9**	**29.3**	**50.3**	**45.4**	**42.0**	**75.6**	1156	**78.7**	**71.1**	121.1	**57.2**

4.4 Ablations

Comparison with Other VLPs. A key property of GenLIP is data efficiency: as shown above, GenLIP pretrained on 8B pairs can surpass baselines pretrained with substantially larger corpora. To further validate this property, we conduct a controlled comparison among a contrastive method (SigLIP), an encoder–decoder generative method (OpenVision2), and our GenLIP under the same pretraining data setting.

Specifically, we train SigLIP, OpenVision2, and GenLIP on the same 2.0B samples from the 1B caption dataset. For GenLIP, we run only the first pretraining stage and evaluate directly at a 384×384 input resolution. For SigLIP and OpenVision2, we pretrain at 224×224 and further conduct a short high-resolution adaptation stage at 384×384 for 0.2B samples. For SigLIP, we implement the vanilla sigmoid contrastive loss without additional tricks from SigLIP2 [67].

We evaluate frozen visual representations of these methods under the same protocol in Table 6. Under the same data budget, GenLIP still achieves strong performance on both Doc&OCR and General VQA tasks. While GenLIP outperforms the baselines on most benchmarks, it trails OpenVision2 on OCRBench by 6.3, which is likely related to the absence of high-resolution adaptation in GenLIP under this controlled setting and the known difficulty of dense-text recognition with low-resolution pretraining.

Overall, this controlled comparison supports that our GenLIP is more data-efficient than both contrastive and prior generative ones.

Comparison with SAIL. GenLIP shares architectural similarities with SAIL [33]. To better understand their differences, we compare SAIL-style Qwen3-0.6B vision encoders with GenLIP under the same 1B-sample pretraining setting in Table 7. Adding gated attention to the Qwen3-initialized SAIL variant brings only a modest overall improvement, while training the same architecture from scratch achieves performance close to GenLIP-So/16. These results suggest that a strong language-model initialization in SAIL is not necessarily beneficial for this vision-encoder pretraining setting. Notably, the parameter count of Qwen3-0.6B is very close to that of GenLIP-so/16, differing only slightly in architecture (e.g., GQA), thus, it is well-suited for comparison.

Gated Attention. In Fig. 4, we plot data scaling curves of GenLIP with and without gated attention, showing consistent advantages of gated attention across data scales. Gated attention improves data efficiency, especially in the low-data

Table 7: Comparison with SAIL-style Qwen3-0.6B vision encoders. "Init" indicates Qwen3-0.6B weight initialization; all methods are trained with the same 1B samples.

Method	Arch	Init	Doc&OCR							General VQA				Caption			ALL AVG
			ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	
SAIL	Qwen3-0.6B	✓	31.6	30.2	39.3	44.1	62.5	25.7	47.3	**41.1**	40.6	**74.6**	1057	74.2	71.9	114.3	53.6
SAIL-g	Qwen3-0.6B	✓	30.2	29.5	39.4	43.9	62.4	26.2	47.8	40.7	41.3	74.2	1141	81.6	75.4	116.9	54.8
SAIL-g	Qwen3-0.6B	✗	32.5	**36.0**	43.6	44.0	63.7	27.4	48.7	40.5	40.6	**74.6**	1131	**81.7**	**77.2**	116.4	56.0
GenLIP	ViT-So/16	✗	**34.6**	34.2	**44.1**	**44.7**	**64.6**	**29.1**	**51.1**	40.7	**41.5**	**74.6**	**1144**	79.9	73.7	**118.4**	**56.3**

Table 8: Comparison of gated attention and register-token variants on GenLIP-So/16 trained with 1B samples from the first pretraining stage.

Method	Doc&OCR							General VQA				Caption			ALL AVG
	ChartQA	OCR-B	DocVQA	TextVQA	AI2D	InfoVQA	SEED-2	VQAv2	GQA	SQA	MME-P	NoCaps	COCO	TextCaps	
1 Register	30.5	32.5	37.2	40.7	63.6	27.4	47.4	36.4	35.6	75.0	**1154**	74.2	71.9	114.3	53.3
4 Registers	34.5	30.8	41.0	43.8	63.8	27.5	50.2	39.7	37.5	**75.5**	1123	76.6	**74.1**	116.9	55.1
Gated Attention	**34.6**	**34.2**	**44.1**	**44.7**	**64.6**	**29.1**	**51.1**	**40.7**	**41.5**	74.6	1144	**79.9**	73.7	**118.4**	**56.3**

regime, where the variant with gated attention achieves higher performance than the one without. It also leads to better convergence and improves the final performance by a notable margin.

In Table 8, we compare gated attention with an alternative approaches for mitigating attention sinks: register tokens. Register tokens introduce additional tokens before the visual sequence to absorb sink behavior in stead of vision tokens. But we find it is better to also keep register tokens in VLMs for better performance, which is different from previous works [15, 58]. Gated attention achieves the better overall average and performs strongest on most Doc&OCR benchmarks than register method with both 1 or 4 register tokens, suggesting that it provides a more effective and flexible solution in our setting.

Discriminative Ability. To assess the discriminative quality of GenLIP's visual representations, we evaluate frozen patch features on ImageNet-1K [16] and ADE20K [82]. Since GenLIP does not use a `[CLS]` token, we adopt attentive probing for classification and a linear probe for semantic segmentation. As shown in Table 9, GenLIP learns transferable discriminative features without class-level or pixel-level visual supervi-

Table 9: Frozen-feature evaluation on ImageNet-1K and ADE20K. For GenLIP models, values marked with (S1) are from the first-stage model; unmarked values use the second-stage model.

Method	Arch	IN-1K	ADE20K
CLIP	L/14	85.1	39.0
SigLIP	So/14	86.7	40.8
SigLIP2	So/16	88.9	45.4
w/o GA	So/16	76.2	-
GenLIP	L/16	83.8(S1) / 82.8	41.0
GenLIP	So/16	84.3(S1) / 83.1	42.8
GenLIP	g/16	85.2(S1) / 83.0	44.5

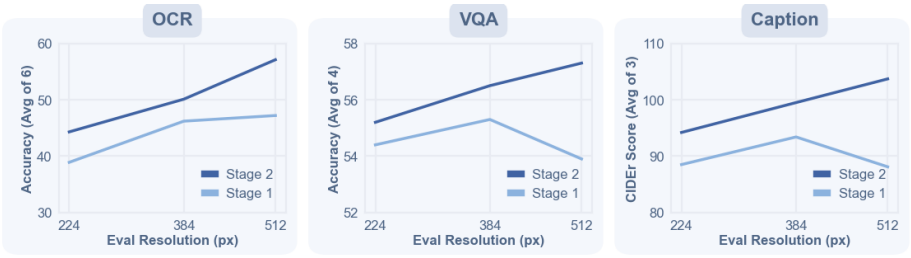

Fig. 5: Validation of Native Aspect Adaptation. Frozen GenLIP-So/16 performance after each pretraining stage under different evaluation resolutions.

sion. Gated attention consistently mitigates the degradation in discriminative representations associated with attention sinks, and the discriminative ability improves with model size. However, since GenLIP is not trained with an explicitly discriminative objective, it still lags behind contrastive methods on these benchmarks.

Native-Aspect-Ratio Adaptation. We evaluate GenLIP from two stages under different resolutions, which validates the effectiveness of the native-aspect-ratio adaptation stage. To test the model's behavior under different input resolutions, we evaluate frozen visual representations of GenLIP (after each stage) across multiple resolutions under the same protocol as in Table 2 (Fig. 5).

5 Conclusions

This work presents GenLIP, a simple generative vision-language pretraining method based on a unified transformer architecture and a single language modeling objective. With a single transformer jointly modeling visual and textual inputs, GenLIP aligns the two modalities through early fusion under one generative objective. Despite its architectural and objective simplicity, GenLIP demonstrates strong data efficiency and scalability for vision-language pretraining, achieving competitive or superior performance across a wide range of multimodal benchmarks with fewer training samples. We hope our exploration of generative vision-language pretraining will inspire future research toward more effective and scalable multimodal learning.

Limitations. Several limitations warrant consideration: (i) our validation experiments are conducted on an academic-scale MLLM setting, LLaVA-NeXT, and the generalizability to cutting-edge MLLMs remains to be verified; (ii) the pretraining dataset is limited to 1.0B scale, and the scaling behavior at even larger volumes remains to be explored; (iii) the reliance on high-quality captions introduces significant data acquisition costs.

Acknowledgments

This work was mainly sponsored by the National Natural Science Foundation of China (No.92470203).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. 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)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
4. An, X., Yang, K., Dai, X., Feng, Z., Deng, J.: Multi-label cluster discrimination for visual representation learning. In: ECCV (2024)
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
7. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
8. Bao, H., Wang, W., Dong, L., Wei, F.: Vl-beit: Generative vision-language pretraining. arXiv preprint arXiv:2206.01127 (2022)
9. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
10. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
11. Chen, Y., Wang, X., Peng, H., Ji, H.: A single transformer for scalable vision-language modeling. Transactions on Machine Learning Research (2024)
12. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
13. 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)
14. Chuang, Y.S., Li, Y., Wang, D., Yeh, C.F., Lyu, K., Raghavendra, R., Glass, J., Huang, L., Weston, J., Zettlemoyer, L., Chen, X., Liu, Z., Xie, S., Yih, W.t., Li, S.W., Xu, H.: MetaCLIP 2: A worldwide scaling recipe. arXiv preprint arXiv:2507.22062 (2025)
15. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
16. 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)

17. Diao, H., Cui, Y., Li, X., Wang, Y., Lu, H., Wang, X.: Unveiling encoder-free vision-language models. Advances in Neural Information Processing Systems **37**, 52545–52567 (2024)
18. Diao, H., Li, M., Wu, S., Dai, L., Wang, X., Deng, H., Lu, L., Lin, D., Liu, Z.: From pixels to words–towards native vision-language primitives at scale. arXiv preprint arXiv:2510.14979 (2025)
19. Diao, H., Li, X., Cui, Y., Wang, Y., Deng, H., Pan, T., Wang, W., Lu, H., Wang, X.: Evev2: Improved baselines for encoder-free vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21014–21025 (2025)
20. 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: 9th International Conference on Learning Representations, ICLR 2021. OpenReview.net (2021)
21. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. Advances in Neural Information Processing Systems **36**, 35544–35575 (2023)
22. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., et al.: Multimodal autoregressive pre-training of large vision encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9641–9654 (2025)
23. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
24. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)
25. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
26. Gu, S., Zhang, J., Zhou, S., Yu, K., Xing, Z., Wang, L., Cao, Z., Jia, J., Zhang, Z., Wang, Y., et al.: Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. CoRR (2024)
27. Huang, Z., Ye, Q., Kang, B., Feng, J., Fan, H.: Classification done right for vision-language pre-training. Advances in Neural Information Processing Systems **37**, 96483–96504 (2024)
28. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
29. Jang, J., Kong, C., Jeon, D., Kim, S., Kwak, N.: Unifying vision-language representation space with single-tower transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 980–988 (2023)
30. 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)

31. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
32. Lai, Z., Zhang, H., Zhang, B., Wu, W., Bai, H., Timofeev, A., Du, X., Gan, Z., Shan, J., Chuah, C.N., et al.: Veclip: Improving clip training via visual-enriched captions. In: European Conference on Computer Vision. pp. 111–127. Springer (2024)
33. Lei, W., Wang, J., Wang, H., Li, X., Liew, J.H., Feng, J., Huang, Z.: The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20758–20769 (October 2025)
34. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. CoRR (2024)
35. Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: SEED-Bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. CoRR (2024)
36. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
37. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
38. Li, X., Zhang, T., Li, Y., Yuan, H., Chen, S., Zhou, Y., Meng, J., Sun, Y., Xu, S., Qi, L., et al.: Denseworld-1m: Towards detailed dense grounded caption in the real world. arXiv preprint arXiv:2506.24102 (2025)
39. Li, X., Liu, Y., Tu, H., Xie, C.: Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3977–3987 (2025)
40. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al.: What if we recaption billions of web images with llama-3? In: International Conference on Machine Learning. PMLR (2024)
41. Li, X., Zhang, F., Diao, H., Wang, Y., Wang, X., Duan, L.: Densefusion-1m: Merging vision experts for comprehensive multimodal perception. Advances in Neural Information Processing Systems **37**, 18535–18556 (2024)
42. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next` (2024)
43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
44. Liu, Y., Li, X., Wang, Z., Zhao, B., Xie, C.: Clips: An enhanced clip framework for learning with synthetic captions. arXiv preprint arXiv:2411.16828 (2024)
45. Liu, Y., Li, X., Zhang, L., Wang, Z., Zheng, Z., Zhou, Y., Xie, C.: OpenVision 2: A family of generative pretrained visual encoders for multimodal learning. arXiv preprint arXiv:2509.01644 (2025)
46. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
47. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)

48. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
49. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
50. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
51. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)
53. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv preprint arXiv:2505.06708 (2025)
54. Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
55. 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)
56. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
57. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European conference on computer vision. pp. 742–758. Springer (2020)
58. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
59. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
60. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14398–14409 (2024)
61. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)

62. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
63. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
64. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
65. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
66. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
67. 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)
68. Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. Advances in Neural Information Processing Systems **36**, 46830–46855 (2023)
69. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
70. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
71. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
72. Xie, Y., Yang, K., An, X., Wu, K., Zhao, Y., Deng, W., Ran, Z., Wang, Y., Feng, Z., Roy, M., Ismail, E., Deng, J.: Region-based cluster discrimination for visual representation learning. In: ICCV (2025)
73. Xing, L., Dong, X., Zang, Y., Cao, Y., Liang, J., Huang, Q., Wang, J., Wu, F., Lin, D.: Caprl: Stimulating dense image caption capabilities via reinforcement learning. arXiv preprint arXiv:2509.22647 (2025)
74. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)
75. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
76. Yang, K., Deng, J., An, X., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Alip: Adaptive language-image pre-training with synthetic caption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2922–2931 (2023)
77. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
78. 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)

79. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: unifying localization and vl understanding. In: 36th Conf. Neural Inf. Process. Syst. NeurIPS (2022)
80. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)
81. Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., Shen, Y.: Dreamlip: Language-image pre-training with long captions. In: European Conference on Computer Vision. pp. 73–90. Springer (2024)
82. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
83. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024)