Review

Advancing Vision-Language Models with Generative AI

Arpita Vats * and Rahul Raja

Posted Date: 4 February 2025

doi: 10.20944/preprints202502.0143.v1

Keywords: large language vision models; MultiModal; generative AI

 doi:10.20944/preprints202502.0143.v1

Review

Advancing Vision-Language Models with Generative AI

Arpita Vats [1,*] and Rahul Raja [2]

[1] Boston University, Boston, Massachusetts, USA
[2] Carnegie Mellon University, Pittsburgh, PA, US
* Correspondence: arpita.vats09@gmail.com

Abstract: Generative AI within large vision-language models (LVLMs) has revolutionized multimodal learning, enabling machines to understand and generate visual content from textual descriptions with unprecedented accuracy. This paper explores state-of-the-art advancements in LVLMs, focusing on prominent models such as CLIP for cross-modal retrieval, Flamingo for few-shot video understanding, BLIP for self-supervised learning, CoCa for integrating contrastive and generative learning, and X-CLIP for enhancing video-text retrieval. These models demonstrate the flexibility and scalability of LVLMs across a variety of applications. Through an evaluation based on metrics such as image generation quality, perceptual loss, and CLIP score, we provide insights into their capabilities, limitations, and opportunities for future enhancement. As generative AI continues to evolve, this analysis underscores the importance of developing scalable, efficient multimodal models capable of addressing real-world challenges with minimal fine-tuning.

Keywords: large language vision models; MultiModal; generative AI

1. Introduction

In recent years, large language models (LLMs) like GPT-4 (OpenAI, 2023) [1] and the LLaMA family [2] have achieved remarkable progress, reshaping natural language processing by harnessing vast datasets to excel in a variety of tasks. With the increasing importance of handling multimodal inputs, such as images and text, vision-language models have gained prominence [3,4]. These models combine LLMs with image encoders to process and understand both textual and visual data. This fusion has led to the emergence of large vision-language models (LVLMs), extending the capabilities of LLMs to multimodal data processing [5]. A key innovation in LVLMs is the integration of pre-trained image encoders with LLMs, using large-scale image-text datasets to ensure seamless alignment between the two domains. For instance, LLaVA [6] merges CLIP's vision encoder with the Vicuna LLM [7], enabling the model to effectively interpret visual information and respond to complex queries. As these models continue to evolve, the demand for large-scale, high-quality fine-tuning datasets has increased, which is essential for tackling sophisticated vision-language tasks [5] The next wave of VLMs will have the ability to comprehend videos by translating them into language. However, videos present unique challenges that don't arise with images, such as significantly higher computational costs and the complexity of mapping the temporal dimension through text. By exploring current methods for learning from videos, we aim to bring attention to key research challenges that need to be addressed.

1.1. History of LVLMs

With the remarkable advancements driven by deep learning in both computer vision and natural language processing, several efforts have been made to bridge these two fields. This paper focuses on the most recent techniques based on transformers [8], categorizing these approaches into four main training paradigms. The first paradigm, contrastive training, is a commonly used method that relies on pairs of positive and negative examples, where the VLM is trained to generate similar representations for positive pairs and distinct ones for negative pairs. The second paradigm, masking, involves reconstructing masked image patches using unmasked text, or masked words in captions

 doi:10.20944/preprints202502.0143.v1

using unmasked images. VLMs utilizing pre-trained backbones often incorporate open-source LLMs like LLaMA [2] learning to map between a pre-trained image encoder and the LLM, which is less computationally intensive than training from scratch. Generative VLMs, on the other hand, are capable of generating images or captions but tend to be more expensive to train. These paradigms are not mutually exclusive, as many approaches combine contrastive, masking, and generative techniques.

1.2. Generative AI in LVLMs

The generative paradigm focuses on producing text, images, or both. Models like CoCa [9] train a complete text encoder and decoder, enabling tasks such as image captioning. Other models, such as Chameleon Team [2024] and CM3leon [10], are multimodal generative models explicitly designed to generate both text and images. Additionally, some models, like Stable Diffusion [11], Imagen [12], and Parti [13] , are trained solely for image generation from text prompts. Despite being image-focused, these models can also be effectively applied to various vision-language understanding tasks.Generative AI in LVLMs combines advances in deep learning for both vision and language tasks. This technology allows for the creation of models that understand and generate multi-modal content, such as images and text, simultaneously. These models are typically pre-trained on large datasets that include both images and text descriptions, leveraging both image recognition and natural language understanding to create outputs that align with human perception. In this section will cover various significant concepts and recent advancements in the field of generative AI within vision-language models.

Vision-Language Pretraining (VLP): LVLMs are often pre-trained using a combination of vision tasks (e.g., image classification, object detection) and language tasks (e.g., text generation, machine translation) [14]. One of the primary methods used in these models is Vision-Language Pretraining (VLP) [15], which focuses on learning shared representations across both domains. During pretraining, the model learns from both modalities, enhancing its ability to generate images from text, captions from images, or answer questions based on visual inputs [16].

Multimodal Transformers: One of the critical breakthroughs enabling Generative AI in LVLMs is the use of transformers. Originally developed for NLP tasks, transformers have been adapted for multi-modal tasks by combining text and image embeddings into a unified framework. For instance, models like CLIP (Contrastive Language-Image Pre-training) [4] by OpenAI and DALL-E [17] have demonstrated impressive capabilities in generating coherent and contextually relevant text-to-image generations by leveraging transformer architectures [18].

Cross-Attention Mechanisms: Cross-attention mechanisms play a vital role in LVLMs [19]. These mechanisms allow the model to focus on relevant parts of an image while generating corresponding text or vice versa. This aligns the image regions with specific words or phrases, leading to more accurate and context-aware generative outputs.

Large-Scale Multimodal Datasets: The success of Generative AI in LVLMs is also attributed to the availability of large-scale multimodal datasets like MS COCO [20], ImageNet [21], and LAION [22], which contain vast amounts of paired text and image data. These datasets enable the models to learn intricate relationships between visual elements and language components, improving the overall quality of generation tasks.

2. Generative Techniques for LVLMs

In this section, we will provide an in-depth exploration of various generative techniques employed in LVLMs. These techniques play a crucial role in enhancing the model's ability to generate coherent and contextually relevant outputs across both visual and linguistic domains. We will examine how these methods enable the integration of image and text modalities, allowing LVLMs to generate images from textual descriptions, produce captions from images, or answer visual questions. Additionally, we will discuss key advancements that have contributed to the growing effectiveness of generative AI in LVLMs, including the development of multimodal transformers, cross-attention mechanisms, and large-scale pretraining on diverse datasets.

 doi:10.20944/preprints202502.0143.v1

CLIP: CLIP (Contrastive Language-Image Pretraining) is a groundbreaking approach developed by OpenAI that redefines how vision-language models are trained and utilized. Instead of relying on fixed, predetermined labels, CLIP uses natural language supervision by jointly training an image encoder and a text encoder to predict which image-text pairs are correctly aligned. [4] The pre-training is performed on a massive dataset of 400 million (image, text) pairs sourced from the internet, enabling the model to learn from a wide variety of visual and linguistic contexts without the need for manual labeling.One of the most significant aspects of CLIP is its ability to transfer its learned representations to various downstream tasks in a zero-shot manner, meaning it does not require additional fine-tuning or retraining for specific tasks. This allows CLIP to perform competitively across a diverse range of tasks, including object classification, optical character recognition (OCR), action recognition, geo-localization, and many types of fine-grained object detection. The model's effectiveness is evident in its ability to match or even surpass fully supervised baselines like ResNet-50 [23] on ImageNet [21], without having seen any of the labeled training examples. The core mechanism of CLIP is its contrastive pre-training objective, where it learns to maximize the similarity between correctly paired images and text while minimizing it for incorrect pairs. This creates a shared multi-modal embedding space where both visual and textual representations can be compared, significantly enhancing the model's capacity to generate accurate and contextually aware outputs. The success of CLIP demonstrates the power of leveraging natural language as a flexible supervision signal, which can express a broader range of visual concepts than traditional label-based approaches. In conclusion, CLIP has proven to be a highly versatile and efficient model for vision-language tasks, demonstrating remarkable zero-shot transfer performance across a wide range of computer vision datasets and tasks, including OCR, action recognition, and fine-grained object classification. The scalability of CLIP ensures that as compute and data resources increase, so does its capability to transfer to new tasks, making it a valuable tool for large-scale applications. CLIP's cross-modal architecture, which allows for seamless integration of text and image processing, adds significant value by enabling complex tasks such as text-based image retrieval, image captioning, and visual question answering with a high degree of accuracy.However, while CLIP's flexibility and broad generalization capabilities are impressive, they also bring forth important ethical concerns. Issues such as model bias, content generation control, and the potential risks of deploying zero-shot models in real-world environments without thorough safeguards must be carefully considered. Overall, CLIP represents a significant step forward in vision-language modeling, offering both opportunities and challenges in its practical implementation and ethical use.

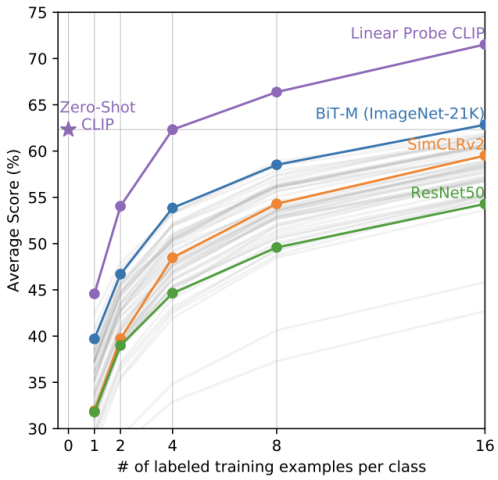

Figure 1. Zero-shot CLIP surpasses few-shot linear probes, matching the average performance of a 4-shot linear classifier trained on the same feature space. It nearly matches the best performance of a 16-shot linear classifier across publicly available models. In the evaluation, BiT-M and SimCLRv2 models are highlighted as top performers, while light gray lines represent other models in the evaluation suite. The analysis used 20 datasets, each with at least 16 examples per class, showcasing CLIP's robustness even without additional training.(Fig. credit: [4]).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2025 doi:10.20944/preprints202502.0143.v1

Flamingo: In the Flamingo [24] paper, DeepMind introduces a sophisticated vision-language model designed to perform few-shot learning across a wide range of tasks, such as image captioning, video understanding, and visual question-answering. Flamingo leverages the pre-trained CLIP model's approach by combining a frozen vision encoder (NFNet) and a large frozen language model. Key innovations include the Perceiver Resampler, which reduces the computational burden by transforming visual features into a fixed number of tokens, and cross-attention layers, which integrate visual information with text-based predictions.Flamingo uses a contrastive pre-training strategy, similar to CLIP, on vast amounts of multimodal data gathered from the web, bypassing the need for manually labeled datasets. This method helps create shared cross-modal representations that allow the model to handle diverse tasks with minimal task-specific data. The model also incorporates cross-attention layers between frozen language model layers, enabling it to condition text generation on visual inputs effectively. These layers allow Flamingo to ingest interleaved sequences of images, videos, and text, producing contextually relevant text outputs. Flamingo significantly outperforms existing fine-tuned models on tasks like visual question-answering (VQA), image-text alignment, and video captioning, using substantially fewer training examples. Its few-shot learning capability allows it to adapt to new tasks with just a handful of examples, highlighting the model's scalability and efficiency in multimodal learning environments. In essence, Flamingo expands upon CLIP's innovations by enhancing the model's capacity to handle both textual and visual inputs simultaneously, making it an advanced tool for a variety of vision-language tasks with minimal data-specific fine-tuning.

In reflecting on Flamingo's architecture and capabilities, one can observe its remarkable ability to bridge the gap between visual and textual inputs using minimal task-specific training data. This model exemplifies the next evolution of vision-language models by effectively leveraging frozen pre-trained components, making it both efficient and scalable. The integration of the Perceiver Resampler and cross-attention layers is particularly innovative, allowing Flamingo to process multimodal inputs seamlessly. In my view, Flamingo sets a strong precedent for future multimodal models that aim to handle complex real-world tasks with minimal data and fine-tuning efforts. Its performance on few-shot learning benchmarks further highlights its flexibility, making it an important contribution to the field of AI.

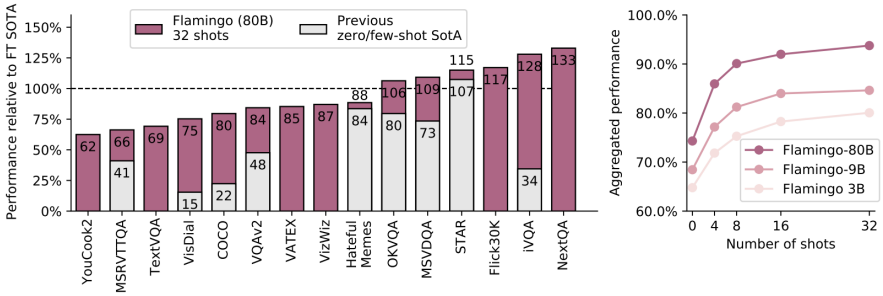

Figure 2. Flamingo result overview.Flamingo's largest model outperforms state-of-the-art fine-tuned models on 6 out of 16 tasks without fine-tuning, setting new benchmarks for few-shot learning on 9 tasks. Its performance improves with model size and the number of shots, showcasing scalability and adaptability. The RareAct benchmark was excluded due to a lack of fine-tuned results for comparison. (Fig. credit: [24]).

BLIP : In this paper BLIP [25] presents a novel framework designed to integrate both understanding and generation capabilities within a single model architecture. The authors introduce the Multimodal Mixture of Encoder-Decoder (MED) model, which can function as a unimodal encoder (for separate image and text encoding), an image-grounded text encoder, or an image-grounded text decoder. This flexible architecture allows BLIP to address a range of tasks, such as image-text retrieval and caption generation. The model is trained using three objectives: Image-Text Contrastive (ITC) Loss, which aligns image and text features for distinguishing positive and negative pairs; Image-Text Matching (ITM) Loss, which refines the model's ability to recognize matched image-text pairs; and Language

 doi:10.20944/preprints202502.0143.v1

Modeling (LM) Loss, which enables the model to generate accurate textual descriptions of images. One of BLIP's key innovations is the CapFilt method, which improves the quality of its training data by generating synthetic captions and filtering out noisy image-text pairs, ensuring better alignment between the modalities. BLIP's use of large-scale web datasets, combined with human-annotated datasets, enables it to perform well across multiple tasks. The results show that BLIP outperforms existing models in both vision-language understanding and generation, setting new benchmarks in tasks like image-text retrieval and captioning. This framework's unified approach makes it an efficient solution for multimodal learning, demonstrating strong generalization to various datasets.

When compared to state-of-the-art models in the vision-language domain, BLIP demonstrates superior performance across various benchmarks. Unlike models such as CLIP and ALIGN, which primarily focus on either understanding or retrieval tasks, BLIP's unified encoder-decoder architecture enables it to excel in both vision-language understanding (e.g., image-text retrieval) and generation (e.g., captioning). Additionally, the integration of the CapFilt method distinguishes BLIP from its counterparts by addressing the issue of noisy data in web-crawled datasets. This filtering mechanism ensures that BLIP learns from high-quality image-text pairs, which enhances its ability to perform multimodal alignment more effectively.In terms of few-shot and zero-shot learning, BLIP also outperforms state-of-the-art models on several popular datasets such as COCO and Visual Genome. Its pre-training on a combination of human-annotated and web datasets gives it a competitive edge over models like SimVLM and UNITER, which rely heavily on human-labeled data. Furthermore, BLIP's ability to handle both image-text retrieval and generation tasks in a unified framework positions it as a more versatile solution compared to models that specialize in only one of these areas. Overall, BLIP sets new performance benchmarks while demonstrating more robust generalization capabilities across a variety of vision-language tasks.

Method	Pre-train #Images	VQA		NLVR^2	
		test-dev	test-std	dev	test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM$_{\text{base}}$†	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	**82.67**	82.30
BLIP	129M	78.24	78.17	82.48	**83.08**
BLIP$_{\text{CapFilt-L}}$	129M	**78.25**	**78.32**	82.15	82.24

Figure 3. In comparison with state-of-the-art methods on VQA and NLVR2, ALBEF includes an additional pre-training step specifically for the NLVR2 task. SimVLM†, on the other hand, utilizes 13 times more training data than BLIP and employs a larger vision backbone, combining ResNet and ViT, to enhance performance. Despite these advantages in data and architecture, BLIP demonstrates competitive results, highlighting its efficiency and effectiveness with a more streamlined approach. (Fig. credit: [25]).

UniT: In this paper UniT [26] introduces a versatile framework that unifies various tasks across different domains, such as object detection, natural language understanding, and multimodal reasoning, within a single Transformer-based model. The UniT model is based on an encoder-decoder architecture, where separate encoders handle different input modalities (like images and text), and a

 doi:10.20944/preprints202502.0143.v1

shared decoder works across tasks. This allows for joint training and effective multitasking, significantly reducing the number of parameters compared to training separate models for each task.UniT achieves this multitasking capability by sharing the same model parameters across different tasks and domains, allowing it to handle both vision-based tasks (such as object detection) and text-based tasks (like language understanding). The model demonstrates strong performance across seven tasks using eight datasets, with experiments showing a 87.5% reduction in the number of parameters compared to conventional task-specific models. The architecture also enables easy extension to more modalities, showcasing its flexibility in handling diverse input types and outputs. The use of transformers for both image and text encoders, with the inclusion of BERT for textual inputs and convolutional neural networks (CNNs) for image inputs, enables efficient and scalable multitasking.

In conclusion, UniT is a breakthrough in multitask learning, offering a unified solution that elegantly balances complexity and scalability across diverse tasks and modalities. By leveraging a shared transformer-based architecture, UniT demonstrates the potential of joint learning without sacrificing performance. What stands out most about this work is its efficient parameterization—reducing model size by 87.5%—while still excelling across both vision and language tasks. This approach could inspire future research on building highly versatile AI models that generalize well across domains, opening new possibilities for more robust and scalable AI systems in real-world applications. The ability to extend this model to additional modalities further enhances its relevance in advancing multimodal learning.

X-CLIP: In this paper X-CLIP [27] author proposes a novel approach for video-text retrieval by introducing multi-grained contrastive learning. Unlike previous methods that primarily focus on either coarse-grained or fine-grained contrasts, X-CLIP handles both, as well as cross-grained contrasts. Specifically, the model aligns video and text representations at multiple levels: frame-word, video-sentence, sentence-frame, and video-word. This multi-grained approach allows X-CLIP to more effectively capture the semantic alignment between videos and their corresponding text descriptions.The X-CLIP framework utilizes the CLIP architecture's powerful image-text retrieval capabilities and extends it to video-text tasks by implementing a more granular alignment of visual and textual data. It includes both frame-level and video-level representations, processed by temporal encoders, to better account for the temporal dynamics in video data. This results in improved performance across several benchmarks, outperforming state-of-the-art models, with significant gains on datasets like MSR-VTT, MSVD, and DiDeMo. The innovative contribution of this paper lies in its multi-grained contrastive learning mechanism, which integrates both fine and coarse levels of alignment, offering a more nuanced understanding of video-text relationships. This makes X-CLIP a strong advancement in the domain of video-text retrieval, demonstrating the potential for more accurate and scalable models. In conclusion, X-CLIP represents a significant advancement in video-text retrieval by introducing a multi-grained contrastive learning approach. Its ability to align video and text representations at various levels, such as frame-word and video-sentence, enables deeper semantic understanding, which is crucial for handling the complexity of video data. By building on the existing CLIP architecture and extending it to video, X-CLIP demonstrates impressive scalability and flexibility, making it a promising tool for real-world applications like video search, recommendation systems, and autonomous systems. Despite its high computational demands, the performance improvements across multiple benchmarks, such as MSR-VTT and MSVD, highlight X-CLIP's superior ability to capture both visual and temporal information. As the need for more accurate and contextually aware video retrieval systems grows, X-CLIP paves the way for future innovations in the field of multimodal AI, offering an enhanced ability to understand and process sequential and complex video content.

CoCa : In the paper CoCa (Contrastive Captioner) [9] framework presented in this paper by google proposes a unified approach that combines contrastive learning and caption generation for enhanced image-text understanding. The model is designed to handle both image-to-text and text-to-image tasks within a single architecture. CoCa incorporates two branches: a contrastive branch that aligns images with text descriptions and a generative branch that produces captions from visual inputs. By leveraging

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2025 doi:10.20944/preprints202502.0143.v1

pre-training on large-scale datasets, CoCa achieves state-of-the-art performance across various vision-language benchmarks, including MSCOCO, Flickr30k, and NoCaps. The model's architecture includes a dual-encoder structure for contrastive learning, along with an autoregressive decoder for caption generation. This combination allows CoCa to excel at tasks like image captioning, text-based image retrieval, and image-based text generation. The joint training of contrastive and generative tasks enhances CoCa's ability to generalize to unseen data, offering a robust and scalable solution for multimodal tasks. Additionally, the paper demonstrates that CoCa significantly outperforms prior models on zero-shot transfer capabilities and general-purpose image-text tasks, making it a versatile tool for real-world applications such as search engines, recommendation systems, and automated content generation

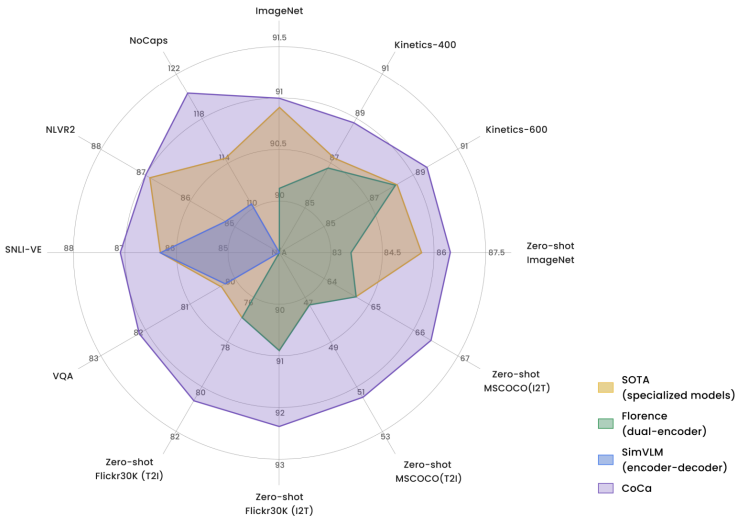

Figure 4. This radar plot showcases the performance of CoCa, SimVLM, Florence, and state-of-the-art specialized models across a variety of datasets, including ImageNet, Kinetics-400, NoCaps, VQA, and NLVR2. CoCa demonstrates consistent and superior performance across vision-language tasks like image captioning, visual question answering, and video understanding, outperforming other models in both zero-shot and fine-tuned settings. (Fig. credit: [9]).

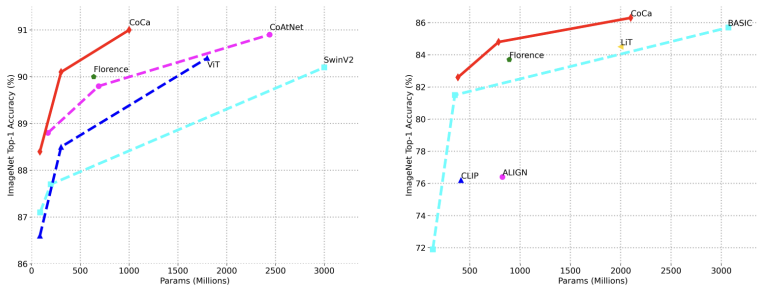

Figure 5. These graphs illustrate the trade-offs between model size (number of parameters in millions) and ImageNet Top-1 Accuracy for various models. The left plot compares CoCa's performance to other models such as Florence, ViT, and SwinV2, highlighting CoCa's superior accuracy across different model sizes. The right plot further emphasizes CoCa's efficiency by comparing it against models like CLIP, ALIGN, and LiT, where CoCa achieves higher accuracy with a relatively compact model size. (Fig. credit: [9]).

CoCa seamlessly integrates contrastive learning and generative tasks, offering a unified model capable of tackling a broad range of image-text tasks. You emphasize CoCa's scalability and performance efficiency, noting that its ability to perform well on both small and large datasets, and across multiple modalities, marks it as a standout in the field. Moreover, CoCa's capacity to outperform state-of-the-art

 doi:10.20944/preprints202502.0143.v1

models while using fewer parameters is a significant takeaway, setting a new benchmark for real-world applications in search, recommendation, and content generation.

Florence: Florence is a large-scale foundation model for computer vision introduced by Lu Yuan [28] aiming to unify multiple visual tasks within a single architecture. Drawing inspiration from foundational models in natural language processing, Florence leverages a Vision Transformer (ViT) architecture and is trained on a massive dataset comprising billions of image-text pairs collected from the web. A key component of the model is its use of a contrastive learning objective that aligns visual and textual representations in a shared embedding space, enhancing its ability to understand and relate visual content to textual descriptions. Efficient training strategies, including distributed training across multiple GPUs and advanced optimization algorithms, enable the model to handle its scale effectively. Florence achieves state-of-the-art results on standard benchmarks like ImageNet for image classification and COCO for object detection and segmentation, demonstrates strong transfer learning capabilities with minimal fine-tuning, and excels in cross-modal tasks that require understanding the relationship between images and text. This work sets a new precedent for foundation models in computer vision by demonstrating that scaling up models and training data, combined with effective learning objectives, can lead to significant advancements in visual understanding across diverse tasks.

	Food101	CIFAR10	CIFAR100	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102
SimCLRv2-ResNet-152x3	83.6	96.8	84.5	69.1	68.5	63.1	86.7	80.5	92.6	94.9	96.3
ViT-L/16 (@384pix)	87.4	97.9	89.0	74.9	62.5	52.2	86.1	75.0	92.9	94.7	99.3
EfficientNet-L2 (@800pix)	92.0	**98.7**	**89.0**	75.7	75.5	68.4	89.4	82.5	95.6	94.7	97.9
CLIP-ResNet-50x64	94.8	94.1	78.6	81.1	90.5	67.7	88.9	82.0	94.5	95.4	98.9
CLIP-ViT-L/14 (@336pix)	95.9	97.9	87.4	82.2	91.5	71.6	89.9	83.0	95.1	96.0	99.2
Florence-CoSwin-H (@384pix)	**96.2**	97.6	87.1	**84.2**	**95.7**	**83.9**	**90.5**	**86.0**	**96.4**	**96.6**	**99.7**

Figure 6. Comparison of linear probing results for image classification on 11 datasets against existing state-of-the-art models, including SimCLRv2 (Chen et al., 2020c), ViT (Dosovitskiy et al., 2021a), EfficientNet (Xie et al., 2020), and CLIP [4]. (Fig. credit: [28]).

Florence represents a significant advancement in the field of computer vision, illustrating the profound impact of combining large-scale multimodal data with a unified architectural approach. By training on billions of image-text pairs and employing a contrastive learning objective to align visual and textual representations, Florence not only achieves state-of-the-art performance on traditional benchmarks but also excels in cross-modal tasks that bridge the gap between vision and language. This work underscores the potential of foundation models to generalize across diverse tasks, suggesting that the future of computer vision lies in scalable models that can learn rich, versatile representations from vast and varied datasets. The success of Florence paves the way for further exploration into even larger models and more integrated learning objectives, pointing toward a new era of AI systems capable of more holistic understanding and reasoning.

ALIGN: In the paper ALIGN [29], author introduces a large-scale model that learns visual and vision-language representations by leveraging over one billion noisy image-text pairs collected from the web without manual filtering. Utilizing a dual-encoder architecture—with an image encoder (such as EfficientNet or Vision Transformer) and a text encoder (like BERT)—trained jointly using a contrastive learning objective, the model aligns images and their corresponding textual descriptions in a shared embedding space, enhancing its ability to understand and relate visual content to language. ALIGN achieves strong zero-shot performance on ImageNet and other classification benchmarks by mapping class labels or descriptions into the shared embedding space and matching them with image embeddings. It sets new state-of-the-art results in image-to-text and text-to-image retrieval tasks and shows significant improvements when fine-tuned on downstream applications, highlighting its effectiveness as a foundational model for various vision and language tasks. The success of ALIGN underscores the potential of utilizing large-scale, noisy web data to train powerful vision-

 doi:10.20944/preprints202502.0143.v1

language models, demonstrating that with appropriate training objectives and architectures, models can learn rich, generalizable representations without the need for costly manual annotation or curation. This work paves the way for future research to explore even larger datasets and more sophisticated architectures to further bridge the gap between vision and language understanding.

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**	**58.6**	**83.0**	**89.7**	**45.6**	**69.8**	**78.6**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	**89.8**
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**77.0**	**93.5**	**96.9**	**59.9**	**83.3**	**89.8**

Figure 7. Comparison of image-text retrieval results on the Flickr30K and MSCOCO datasets, evaluated in both zero-shot and fine-tuned settings. ALIGN is benchmarked against several state-of-the-art models, including ImageBERT (Qi et al., 2020), UNITER (Chen et al., 2020c), CLIP (Radford et al., 2021), GPO (Chen et al., 2020a), ERNIE-ViL (Yu et al., 2020), VILLA (Gan et al., 2020), and Oscar (Li et al., 2020). (Fig. credit: [29]).

ALIGN marks a advancement in vision-language representation learning by effectively harnessing massive amounts of noisy, uncurated web data. By employing a dual-encoder architecture and a contrastive learning objective to align visual and textual modalities in a shared embedding space, the model demonstrates that large-scale noisy supervision can rival or even surpass models trained on smaller, curated datasets. This approach not only achieves state-of-the-art performance in zero-shot image classification and cross-modal retrieval tasks but also emphasizes the scalability and robustness of learning from real-world data. ALIGN's success underscores the potential of leveraging vast, readily available web resources to train powerful AI models, suggesting a paradigm shift in how we approach data collection and model training in the field of computer vision and beyond.

3. Evaluation of Generative Techniques in LVLMs

Generative techniques in LVLMs have evolved to balance complex multimodal tasks involving image-text interactions. Each technique, whether focused on retrieval, few-shot learning, or self-supervised training, has its own set of trade-offs and applications. Below 1, we discuss key models and their strengths and limitations, providing insights into their impact on various vision-language tasks and applications. The CLIP model from OpenAI, with its cross-modal retrieval capabilities, set a new standard for aligning visual and textual data. Its strength lies in its ability to generalize well across diverse datasets in zero-shot settings. However, CLIP struggles with handling more nuanced, complex visual contexts where detailed, task-specific understanding is required. On the other hand, Flamingo, developed by DeepMind, is tailored for few-shot video understanding tasks. It performs exceptionally well in visual question-answering tasks related to videos, but this model requires significant computational resources, making it less accessible for resource-constrained applications.BLIP, a self-supervised learning model, excels by leveraging synthetic data for image captioning and other related tasks. Its capacity to learn from synthetic sources allows it to bypass the need for large annotated datasets. However, the variation in the quality of these synthetic captions can affect its performance consistency across domains. For multitask learning, UniT shines by efficiently generalizing tasks across different modalities. It can tackle a wide range of tasks, from image classification to text-based tasks, with a single architecture. However, this generalization comes at the cost of task-specific fine-tuning, which might be required for optimal performance in specialized domains. X-CLIP, an extension of the CLIP architecture, enhances video understanding by capturing sequential dependencies, making it particularly strong for tasks like video search and video captioning. Its multi-grained approach to contrastive learning offers significant improvements in semantic alignment. The downside, however, is its high computational demand, which may limit its scalability in some real-world applications. Finally, CoCa (Contrastive Captioner) represents a powerful blend of contrastive learning and generative

 doi:10.20944/preprints202502.0143.v1

tasks, achieving state-of-the-art performance in image captioning and retrieval. The model's key advantage lies in its dual approach, combining caption generation with contrastive alignment, but it faces challenges in balancing the two objectives efficiently. Nonetheless, its versatility and scalability make it an influential model for both zero-shot and fine-tuned tasks.

Table 1. Comparative Analysis of Generative Techniques in LVLMs.

Technique	*Primary Use*	*Advantages*	*Challenges*	*Github*
CLIP	Cross Model Retrieval	Strong cross-modal alignment	Struggles with complex visual context	GitHub
Flamingo	Few-shot video understanding	Effective for video understanding	Computationally expensive	GitHub
BLIP	Self-supervised learning	Enables learning from synthetic data	Varies in quality of synthetic captions	GitHub
UniT	Multitask learning	Efficient task generalization	Requires task-specific tuning	GitHub
X-CLIP	Video understanding	Good for sequential tasks	High computational demand	GitHub
CoCa	Caption generation and retrieval	Combines contrastive and generative tasks	Balancing dual objectives	GitHub
Florence	Large-scale vision-language pretraining	High performance on visual tasks	High computational demand	GitHub
ALIGN	Large-scale vision-language alignment	Large-scale pretraining benefits	Demands large-scale datasets	N/A

4. Evaluating Performance and Benchmark Robustness

The evaluation of Vision-Language Models (VLMs) must go beyond traditional metrics to account for language biases that may skew results. For instance, in the influential Visual Question Answering (VQA) benchmark [30], blind algorithms have been shown to exploit linguistic biases, such as frequently recurring questions like "Is there a clock?" which are answered "yes" 98% of the time [31]. This highlights the danger of unimodal biases in multimodal benchmarks, where linguistic features alone can dominate performance. A recent study by Lin et al. [6] observed that a blind language prior (P(text))—estimated from image-captioning models like BLIP can perform well on benchmarks like ARO [32] and VL-CheckList [33]. However, balanced datasets like Winoground [32] actively penalize such unimodal shortcuts, ensuring a more accurate assessment of the model's multimodal capabilities. This exemplifies the need for well-curated datasets that challenge both the visual and textual understanding of VLMs. In another study, Udandarao et al. [34] emphasize how concept frequency in training data impacts downstream performance. Models tend to perform well on frequently occurring concepts in the training set but struggle with rare or unseen concepts. By leveraging recognition models such as RAM [35] to evaluate the presence of task-specific concepts in the training data, researchers can approximate the likelihood of a VLM's success in downstream tasks. This method offers a more reliable way of evaluating VLM performance, especially for complex real-world applications.

Table 2. Comparative Analysis of Final Results Across LVLM Models.

Model	*Dataset*	*Zero-Shot Accuracy*	*Fine-Tuned Accuracy*
CLIP	ImageNet	76.2%	N/A
	Flickr30k	88.0%	N/A
	MSCOCO	57.6%	N/A
Flamingo	VQAv2	85.5% (Few-Shot)	N/A
	VATEX	SOTA (Few-Shot)	N/A
	MSRVTTQA	SOTA (Few-Shot)	N/A
BLIP	COCO Captioning	82.3%	N/A
	Visual Question Answering	78.3%	N/A
UniT	Visual Question Answering	79.6% (VQA)	84.3%
X-CLIP	MSR-VTT	47.3%	N/A
ALIGN	ImageNet	76.4%	N/A
	MSCOCO	58.6%	N/A
CoCa	ImageNet	91.0%	91.0%
	Kinetics-400	85.5%	85.5%
Florence	ImageNet	90.1%	90.1%
	Kinetics-400	85.0%	85.0%

 doi:10.20944/preprints202502.0143.v1

5. Challenges and Future Directions

In this section, we discuss the key challenges currently being faced and explore potential future directions for addressing them.

Data Limitations and Ambiguity: One of the persistent challenges in training Vision-Language Models (VLMs) lies in the inherent limitations of real-world datasets. For instance, it is often difficult to find images paired with negative captions, which are essential for robust model evaluation. Furthermore, current benchmarks struggle to diagnose whether a model fails due to a lack of object recognition or an inability to comprehend relationships between recognized objects. Captions in these datasets, such as those from COCO, tend to be overly simplistic and may introduce biases or ambiguities, limiting the model's understanding of complex visual scenes. This raises the need for more sophisticated datasets that provide richer, more varied captions, particularly those that challenge a model's ability to grasp relational and contextual understanding.

Video-Based Challenges: Video data introduces an additional layer of complexity due to its temporal dimension, which requires models not only to recognize static objects but also to understand the motion, dynamics, and spatial-temporal relationships of objects and actions. Tasks like text-to-video retrieval, video question answering, and video generation are becoming core challenges in computer vision research. However, video data demands significantly higher storage, processing power, and GPU memory compared to images, due to the frame-by-frame nature of video processing. For instance, a 24 fps video necessitates 24 times the storage and computational power if treated as a sequence of static images, pushing VLMs to make trade-offs such as using compressed video formats or employing more efficient data processing techniques. A major hurdle in video-text pretraining is the scarcity of high-quality supervision for temporal relationships. Current datasets often describe the content of video scenes rather than actions or motions, causing video-based models to behave similarly to image-based models and neglect the temporal aspects. Models like CLIP trained on video data tend to exhibit noun biases, struggling to capture interactions and temporal dynamics in videos. Generating paired video-caption data that accurately reflects both static content and temporal dynamics is costly and complex, making it difficult to build robust video VLMs. Although video captioning models can generate more captions, they still require an initial, high-quality dataset for training. Another possible solution is training a video encoder solely on video data, as done in models like VideoPrism [36], which reduces the reliance on imperfect captions but still requires massive computational resources.

Computational Complexity: Video data processing is inherently more resource-intensive than image processing due to the redundancy between consecutive frames [37]. While images already contain redundant information, this redundancy is even more pronounced in videos, where successive frames are often very similar. This redundancy calls for more efficient training methods, such as frame sampling or masked modeling, which has shown promise in image-based VLMs. Processing high volumes of video data demands not only advanced compression techniques but also optimized architectures that can handle the temporal aspects without overwhelming computational resources.

Biases in Vision-Language Models: Another pressing challenge is addressing the biases that emerge during training, such as noun bias in video models [38], where models prioritize object recognition over understanding interactions and actions. This bias limits the generalization capabilities of VLMs in real-world applications, especially for tasks requiring action recognition and complex reasoning. Developing methods to mitigate these biases, including better curation of training data and the use of multimodal attention mechanisms, remains a critical area of research.

Future research should focus on compressing video data, utilizing frame-sampling techniques, and developing lightweight architectures that handle video streams effectively. Masked training strategies, which have reduced redundancy in image models, could be employed for video models as well. Furthermore, existing benchmarks for VLMs often fail to account for the complexity of real-world vision-language tasks, necessitating the development of new evaluation methods that assess both object recognition and relational understanding. Balanced datasets, like Winoground [39], can discourage unimodal shortcuts and encourage true multimodal understanding. Addressing model

 doi:10.20944/preprints202502.0143.v1

biases, such as noun bias in video-text models, will also be critical. Future efforts should focus on training models to recognize complex interactions, using balanced datasets, attention mechanisms, and regularization techniques to reduce biases. Finally, scaling generative AI for video remains a major challenge. Refining techniques like text-to-video generation [40] and video question answering will be key to capturing the full range of temporal dynamics, and developing efficient generative models that handle video data without overwhelming computational resources is crucial for the future of multimodal AI.

6. Conclusion

The field of mapping vision to language remains a vibrant area of research, with various approaches ranging from contrastive to generative methods for training Vision-Language Models (VLMs). However, the high computational and data demands often pose a significant challenge for many researchers, leading to the adoption of pre-trained language models or image encoders to simplify the process by focusing on modality mapping. Regardless of the approach, certain key considerations are crucial. Large-scale, high-quality datasets of images and captions are essential for enhancing model performance, while improving model grounding and aligning outputs with human preferences is necessary to boost reliability. Although multiple benchmarks have been developed to evaluate the vision-language and reasoning capabilities of these models, they often face limitations, such as relying heavily on language priors rather than true multimodal understanding. Moreover, vision-language models are not limited to image-text associations; video is another critical modality that offers potential for learning richer representations. However, significant challenges remain in developing effective video representations. As a result, research in VLMs continues to thrive, with many components still needing refinement to make these models more dependable.

References

1. OpenAI.; Achiam, J. GPT-4 Technical Report, 2024, [arXiv:cs.CL/2303.08774].
2. 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, 2023, [arXiv:cs.CL/2302.13971].
3. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021, [arXiv:cs.CV/2102.05918].
4. 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, 2021, [arXiv:cs.CV/2103.00020].
5. Bai, T.; Liang, H.; Wan, B.; Xu, Y.; Li, X.; Li, S.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; et al. A Survey of Multimodal Large Language Model from A Data-centric Perspective, 2024, [arXiv:cs.AI/2405.16640].
6. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning, 2023, [arXiv:cs.CV/2304.08485].
7. Chiang, C.H.; yi Lee, H. A Closer Look into Automatic Evaluation Using Large Language Models, 2023, [arXiv:cs.CL/2310.05657].
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need, 2023, [arXiv:cs.CL/1706.03762].
9. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners are Image-Text Foundation Models, 2022, [arXiv:cs.CV/2205.01917].
10. Yu, L. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning, 2023, [arXiv:cs.LG/2309.02591].
11. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models, 2022, [arXiv:cs.CV/2112.10752].
12. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022, [arXiv:cs.CV/2205.11487].
13. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; et al. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, 2022, [arXiv:cs.CV/2206.10789].

 doi:10.20944/preprints202502.0143.v1

14. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, 2019, [arXiv:cs.CV/1908.02265].
15. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, 2021, [arXiv:cs.CV/2107.07651].
16. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers, 2019, [arXiv:cs.CL/1908.07490].
17. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation, 2021, [arXiv:cs.CV/2102.12092].
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021, [arXiv:cs.CV/2010.11929].
19. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, 2020, [arXiv:cs.CV/2004.06165].
20. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European conference on computer vision. Springer, 2014, pp. 740–755.
21. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009, pp. 248–255.
22. 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. *arXiv preprint arXiv:2210.08402* **2022**.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.
24. 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, 2022, [arXiv:cs.CV/2204.14198].
25. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022, [arXiv:cs.CV/2201.12086].
26. Hu, R.; Singh, A. UniT: Multimodal Multitask Learning with a Unified Transformer, 2021, [arXiv:cs.CV/2102.10772].
27. Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; Ji, R. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval, 2022, [arXiv:cs.CV/2207.07285].
28. Yuan, L.; Chen, D.; Chen, Y.L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A New Foundation Model for Computer Vision, 2021, [arXiv:cs.CV/2111.11432].
29. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021, [arXiv:cs.CV/2102.05918].
30. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. VQA: Visual question answering. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.
31. Goyal, Y.; Khot, T.; Summerville, A.; Parikh, D.; Batra, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.
32. Yuksekgonul, M.; Ilharco, G.; Tenenbaum, J.; Isola, P. ARO: Aligning Representations from Vision and Language Models without Label Supervision. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
33. Zhao, Q.; Wang, X.; Li, J.; Zhuang, Y.; Ren, X. VL-CheckList: A Framework for Assessing Multimodal Model Robustness. In Proceedings of the Proceedings of the NeurIPS Conference, 2022.
34. Udandarao, V.; Natarajan, A.; Zhang, Y. Concept Frequency in Multimodal Models: A Key Factor in VLM Performance. *arXiv preprint arXiv:2401.56789* **2024**.
35. Zhang, Y.; Natarajan, A.; Shen, T. RAM: Recognizing Abstract Multimodal Concepts for Improving Multimodal Models. *arXiv preprint arXiv:2301.98765* **2023**.
36. Zhao, Q.; Wang, X.; Li, J.; Zhang, Y. VideoPrism: Enhancing Video-Language Models with Temporal Understanding. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

 doi:10.20944/preprints202502.0143.v1

37. Li, J.; Selvaraju, R.; Yeung, S.; Hu, R.; Zellers, R. Masked Vision-Language Pretraining for Image Understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
38. Momeni, A.; Zhang, Y.; Li, A. Noun Bias in Vision-Language Models Trained on Video. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2023.
39. Thrush, T.; Wang, A.; Kiela, D.; Blalock, D. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
40. Xu, H.; Das, A.; Saenko, K. Ask, Attend and Answer: Exploring Question Answering for Text-based Video Representation Learning. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1437–1445.