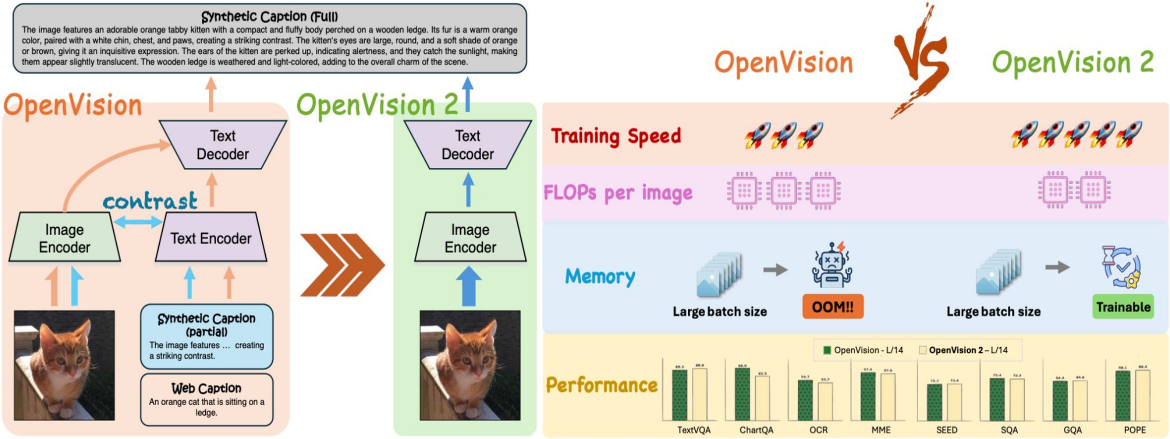
OpenVision 2 : A Family of Generative Pretrained Visual Encoders for Multimodal Learning

Yanqing Liu[1] Xianhang Li[1] Letian Zhang[1] Zirui Wang[2] Zeyu Zheng[3] Yuyin Zhou[1] Cihang Xie[1]
[1]University of California Santa Cruz [2]Apple [3]University of California Berkeley

Project Page: https://ucsc-vlaa.github.io/OpenVision2
Model Training: https://github.com/UCSC-VLAA/OpenVision

Contributions

- Challenges the belief that **CLIP-style contrastive learning is essential** for vision encoders.
- OpenVision 2 shows **a caption-only generative objective can match multimodal performance**.
- Approach **reduces computation and memory costs** compared to contrastive methods.
- Full training suite and pretrained checkpoints of **OpenVision 2 are publicly released**.



OpenVision

Model	Data	Training	Evaluation	Model Num.	Training Time
OpenAI's CLIP	Closed	Closed	Open	4	Short Long
Google's SigLIP	Closed	Open	Open	10	
OpenVision	***Open***	***Open***	***Open***	***>25***	

Fully-Open Vision Encoders

- Open release of datasets, training recipes, and model checkpoints for transparency and reproducibility.

Wide Range of Model Scales

- A family of encoders ranging from *Tiny* (~5.9 million params) to *Huge* (~632.1 million).
- Flexibility for deployment across a spectrum from edge devices to big compute servers.

Superior Multimodal Performance

- Matches or exceeds performance of proprietary encoders (e.g. OpenAI's CLIP, Google's SigLIP) on several multimodal benchmarks, particularly in frameworks like LLaVA-1.5 and Open-LLaVA-Next.

Efficiency: Progressive & Resolution Training

- Use of progressive resolution training (start with lower resolution images, move to higher) to speed up training and save compute.
- Significant reductions in training time and memory usage in comparison with existing large models and proprietary CLIP models.

OpenVision

Efficiency (CLIPA @UCSC)

- OpenVision adopts this two-stage curriculum by training on low-resolution images(84^2)and conducting a fine-tuning at full resolution(224^2).

Data Quality (Recap @UCSC)

- a LLaMA-3-powered LLaVA model recaptions the entire DataComp-1B collection; this high-quality synthetic set serves as the training corpus for OpenVision.

Optimization (CLIPS @UCSC)

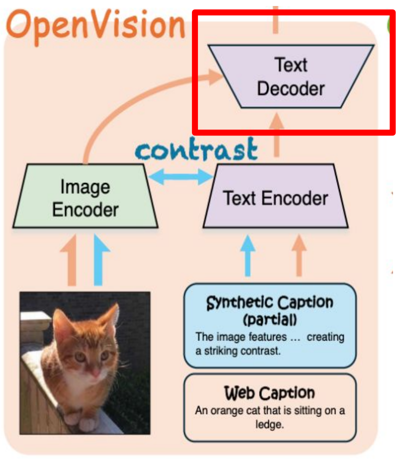
- To better leverage synthetic captions, CLIPS introduces two additional objectives: (i) **a dual contrastive loss** that pairs each image with both web-crawled and generated captions, and (ii) **a caption loss** that asks the model to predict the synthetic caption given the image and its web caption. OpenVision integrates both losses to enhance training.

OpenVision2

OpenVision Challenges

- the text encoder must process **two captions per image for the dual contrastive objective**
- **an additional text decoder is required** to autoregressively predict the synthetic caption.

→ Together, these two components substantially increase FLOPs and GPU memory in training.



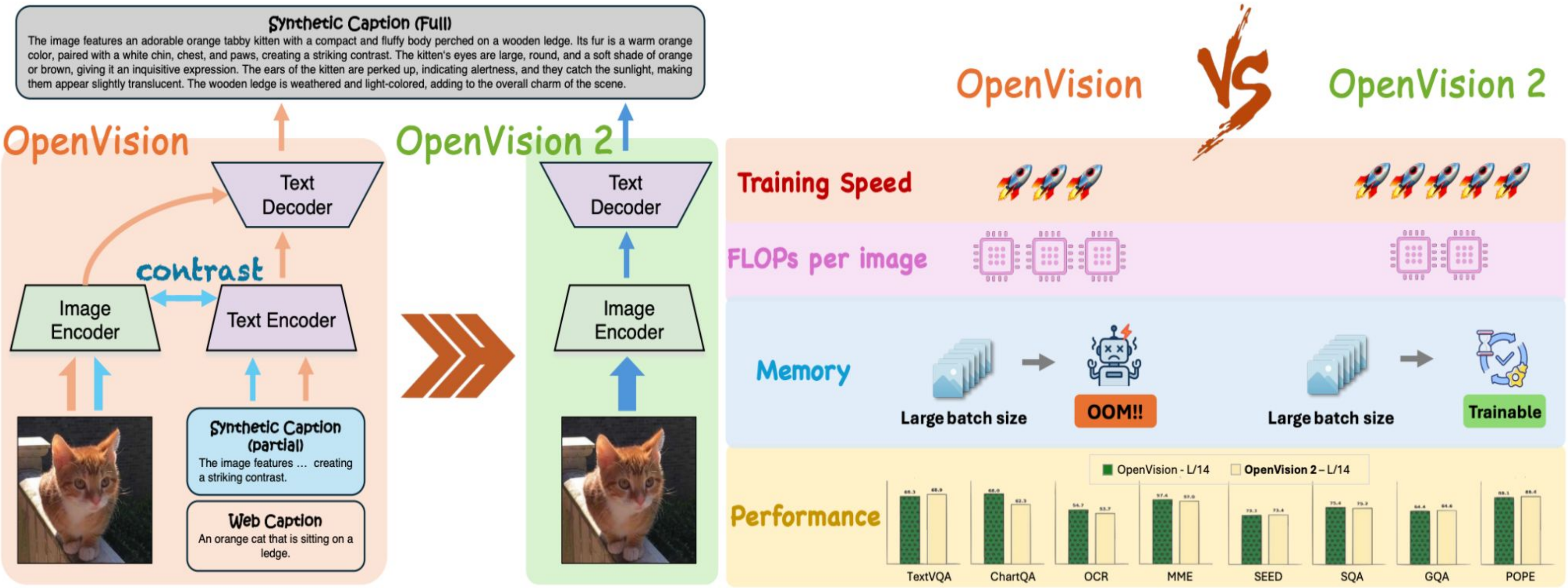
OpenVision 2 approaches

- Discarding the text encoder and contrastive loss.
- Simplifying training to:
    1. Vision encoder → visual tokens.
    2. Text decoder → synthetic caption.

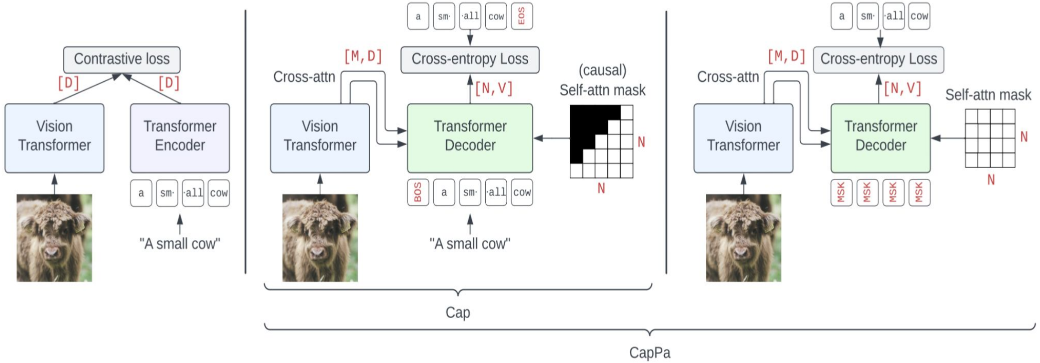
→ This makes pretraining **purely generative**, aligning better with downstream fine-tuning (e.g., LLaVA).

→ Efficiency tweak: randomly mask ~⅔ of visual tokens, which still allows good captioning while reducing computation.

OpenVision2



CapPa



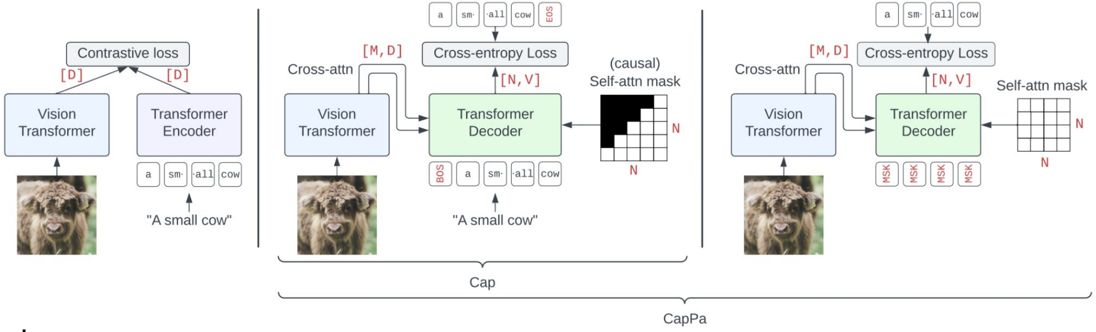
Background

- **Contrastive Pretraining Dominance**
- **Generative Captioning Considered Inferior**
- **Lack of Fair Comparisons**

Trade-offs

- Zero-shot classification: **contrastive wins in many standard benchmarks**.
- Fine-grained tasks, compositionality, ordering, relations: contrastive models tend to ignore word order or relational structure and treat text more like a “bag of words.” **Captioning models are potentially better at capturing these finer structures**.
- Efficiency / inference cost: captioning models (encoder-decoder) require decoding (autoregressive or parallel), which is costlier in some settings compared to just encoding text/image separately (as in CLIP).

CapPa



Architecture

- Vision Transformer (ViT) as the image encoder.
- Standard Transformer decoder that takes the encoder's output via cross-attention to generate captions.
- The decoder has fewer layers (half the depth) than the encoder in their setups, but matches width & attention heads.

Captioner (Cap) Variant

- Pure image captioning: autoregressive decoding (teacher forcing) predicting next token given previous text tokens + image encoding.

Parallel Prediction / Mixed Mode ("CapPa")

- CapPa uses a *parallel prediction* mode for a fraction of training data: the decoder's input text tokens are masked (all [MASK]), attention mask is changed so there's no causal masking. The decoder must predict all tokens at once (positions matter) given only the image (not previous text tokens).

Difference from CapPa

Higher-quality captions

- Uses *ReCap-DataComp-1B* (Llama-3 recaptioned dataset) with improved captioning strategy → produces longer, more grounded captions for stronger generative supervision.

Fusion simplification

- Replaces CapPa's cross-attention with simple concatenation of visual tokens in the text decoder; randomly drops tokens during training to regularize and reduce cost.

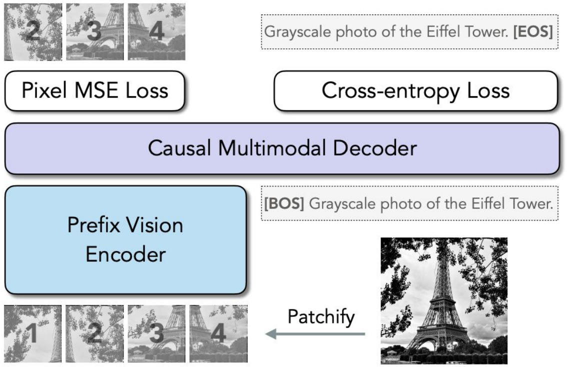
Scale & evaluation

- Scales vision encoder to **1.01B parameters** trained on **12.8B image–caption pairs**; evaluates on advanced benchmarks (MME, ChartQA), beyond classification/QA.

Decoding strategy

- Uses **standard autoregressive decoding** only, instead of CapPa's hybrid approach.

AIMv2



Background

- 대표적인 modal align 방법은 generative vs discriminative(contrastive)
- generative 직관적인 사전학습 방법, 그러나 높은 capacity
- discriminative 방법은 parameter efficient, 그러나 학습이 까다로움

  → generative 사전 학습의 간단함과 확장성 그리고 discriminative 방법의 parameter-efficient 방법을 고려

Architecture

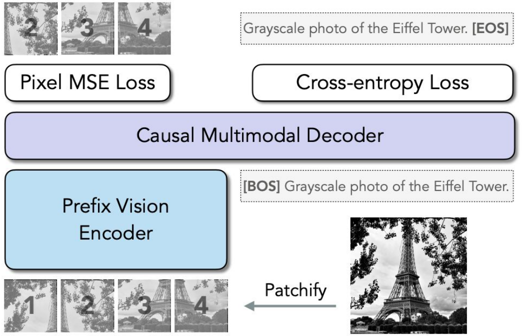
- 구성: vision encoder + multimodal decoder (autoregressive로 다음 이미지 패치와 텍스트 토큰을 예측할 수 있도록)
- ViT architecture (300M~3B)
- Prefix attention (to facilitate the use of bidirectional attention)
- SwiGLU + RMSNorm
- Multimodal decoder
  - The outputs of the decoder are processed through two separate linear heads to predict the next token in each modality respectively. (image token head, text token head)

https://docs.google.com/document/d/1XIA8pPkgGNwQzh_9l061TX2t5keT4RMqnLih1gpZHCA/edit?usp=sharing

AIMv2

Contributions

- 쉽고 직관적인 사전학습 방법 (이미지 패치와 텍스트 토큰을 같이 auto-regessively 예측하는 causal multimodal encoder 사용, 이때 contrastive 방법과 달리 배치 사이즈, 배치 간 고려는 하지 않기 때문에 학습하기 쉬움)
- 확장성 (다양한 모달 지원)
- localization, grounding, classification포함한 비전 벤치마크, 멀티-모달 벤치마크를 포함 우수한 성능
- 멀티모달 이해에서도 CLIP SigLIP과 같은 sota모델보다 우수함.



모델	특징 및 강점	단점
AIMv2	멀티모달 및 이미지 이해에서 높은 성능, 다양한 해상도 지원	학습 및 사용을 위해 고사양 필요
CLIP	텍스트-이미지 align 작업에서 높은 성능	멀티모달 확장성 제한적
DINOv2	객체탐지에서 우수성능	멀티모달성능은제한적

https://docs.google.com/document/d/1xsgDB5OzA4wqtveUX65UbuGP6YJh3GnAdA8S4bbNvLU/edit?usp=sharing

Difference from AIMv2

Training signal

- AIMv2 → combines image patch reconstruction + text generation.
- OpenVision 2 → caption-only supervision (no image reconstruction).

Token masking

- OpenVision 2 masks ~⅔ of visual tokens → improves efficiency & performance.

Data composition

- AIMv2 → mix of human (67%) + synthetic (33%) captions.
- OpenVision 2 → fully synthetic captions from ReCap-DataComp-1B (richer & consistent).

Vision encoder

- AIMv2 → prefixViT with special attention mask.
- OpenVision 2 → standard ViT backbone (simple & efficient).

Results

Under LLaVA 1.5 Framework

Method	Vision Encoder	Params	# Res.	Text VQA	Chart QA	OCR.	MME	SEED	SQA	GQA	POPE
OpenAI-CLIP [44]	L/14	304M	224	56.1	13.2	177	1443/306	66.0	73.4	60.8	85.0
LAION-2B-CLIP [19]	L/14	304M	224	54.2	12.8	165	1434/298	65.5	76.0	59.0	84.5
DataComp-1B-CLIP [16]	L/14	304M	224	53.0	12.3	131	1382/312	62.4	74.2	57.8	83.0
DFN-2B-CLIP [12]	L/14	304M	224	53.2	12.4	246	1447/306	65.6	76.3	59.1	85.0
MetaCLIP-5B [59]	L/14	304M	224	55.6	12.8	313	1552/315	67.4	78.0	61.3	85.4
OpenVision [30]	L/14	304M	224	57.7	13.9	315	1487/317	69.5	73.6	62.9	86.4
OpenVision 2	**L/14**	**304M**	**224**	**59.0**	**13.7**	**327**	**1460/312**	**69.3**	**76.5**	**62.6**	**87.1**
OpenAI-CLIP [44]	L/14	304M	336	59.1	13.8	201	1475/288	67.5	73.1	61.1	85.7
OpenVision [30]	L/14	304M	336	61.2	15.7	339	1525/315	70.5	75.1	63.7	87.2
OpenVision 2	**L/14**	**304M**	**336**	**63.0**	**14.5**	**357**	**1486/321**	**70.1**	**77.5**	**63.0**	**87.7**
SigLIP [62]	SoViT-400M/14	400M	384	62.6	14.5	338	1481/347	69.4	76.7	63.3	87.0
OpenVision [30]	SoViT-400M/14	400M	384	62.4	16.1	357	1493/320	70.4	72.4	63.8	88.0
OpenVision 2	**SoViT-400M/14**	**400M**	**384**	**64.3**	**15.0**	**387**	**1472/310**	**70.7**	**74.9**	**63.5**	**87.5**
OpenVision 2	**H/14**	**632M**	**224**	**60.2**	**13.5**	**340**	**1470/305**	**69.3**	**75.4**	**62.5**	**87.2**
OpenVision 2	**H/14**	**632M**	**336**	**63.4**	**16.3**	**391**	**1470/311**	**70.6**	**76.4**	**63.1**	**88.4**
OpenVision 2	**H/14**	**632M**	**448**	**65.6**	**18.1**	**416**	**1499/331**	**70.6**	**75.6**	**63.1**	**88.7**
OpenVision 2	**g/14**	**1.01B**	**224**	**60.2**	**13.7**	**338**	**1469/290**	**69.3**	**75.0**	**62.6**	**86.9**

Results

Under Open-LLaVA next Framework

Method	Vision Encoder	Params	# Res.	Text VQA	Chart QA	OCR.	MME	SEED	SQA	GQA	POPE
OpenAI-CLIP [44]	L/14	304M	224	62.8	60.7	459	1600/334	70.6	75.0	62.8	86.9
LAION-2B-CLIP [19]	L/14	304M	224	59.4	50.8	396	1533/323	70.0	72.9	62.7	86.4
DataComp-1B-CLIP [16]	L/14	304M	224	58.1	48.5	373	1524/348	70.2	75.6	62.3	86.2
DFN-2B-CLIP [12]	L/14	304M	224	57.0	42.7	303	1486/328	68.3	70.6	61.7	86.0
MetaCLIP-5B [59]	L/14	304M	224	63.0	62.9	493	1590/335	72.3	77.1	64.0	86.8
OpenVision	L/14	304M	224	65.7	61.5	503	1567/332	73.1	73.1	64.7	87.8
OpenVision 2	**L/14**	**304M**	**224**	**66.1**	**60.4**	**501**	**1577/297**	**73.1**	**68.4**	**64.6**	**87.6**
OpenAI-CLIP [44]	L/14	304M	336	69.4	70.0	535	1591/351	73.3	76.9	64.5	87.6
OpenVision	L/14	304M	336	68.3	68.0	547	1520/310	73.3	75.4	64.4	88.1
OpenVision 2	**L/14**	**304M**	**336**	**68.9**	**62.3**	**537**	**1585/278**	**73.4**	**75.2**	**64.6**	**88.4**
SigLIP [62]	SoViT-400M/14	400M	384	68.2	61.3	494	1539/325	72.9	74.7	62.9	86.8
OpenVision	SoViT-400M/14	400M	384	67.4	63.1	540	1500/353	72.2	73.5	63.4	87.8
OpenVision 2	**SoViT-400M/14**	**400M**	**384**	**69.0**	**63.4**	**549**	**1521/319**	**72.2**	**72.7**	**63.1**	**87.7**
OpenVision 2	**H/14**	**632M**	**224**	**66.4**	**60.2**	**514**	**1597/314**	**73.3**	**76.2**	**64.7**	**88.4**
OpenVision 2	**H/14**	**632M**	**336**	**69.9**	**64.8**	**573**	**1572/337**	**73.8**	**74.5**	**64.4**	**87.8**
OpenVision 2	**H/14**	**632M**	**448**	**71.9**	**64.9**	**590**	**1542/324**	**74.1**	**75.6**	**64.4**	**88.8**
OpenVision 2	**g/14**	**1.01B**	**224**	**67.3**	**62.4**	**514**	**1558/323**	**73.4**	**74.4**	**64.7**	**88.0**

Results

Model	Backbone	Resolution	v4-512 Hours	FLOPs / Image
OpenVision [30]	L/14	224	83	271.75
OpenVision 2	**L/14**	**224**	**57**	**208.90**
OpenVision [30]	SoViT-400M/14	384	241	1636.75
OpenVision 2	**SoViT-400M/14**	**384**	**121**	**1017.74**

OpenVision 2 achieves faster training and lower computational cost across model sizes.

Model	Resolution	Batch Size	Peak Memory (GB)
OpenVision [30] (L/14)	224	2k	24.5
	224	4k	OOM
OpenVision 2 (L/14)	**224**	**2k**	**13.8**
	224	**4k**	**22.1**
	224	**8k**	**28.4**
OpenVision [30] (SoViT-400M/14)	384	512	27.4
	384	1k	OOM
OpenVision 2 (SoViT-400M/14)	**384**	**512**	**14.5**
	384	**1k**	**28.8**

OpenVision 2 achieves faster training and lower computational cost across model sizes.

Results

Caption Type	Text VQA	Chart QA	OCR.	MME	SEED	SQA	GQA	POPE
Alt-text	51.8	12.3	238	1306/293	58.6	75.3	55.4	82.2
ReCap-DataComp-1B	56.9	12.9	291	1426/293	67.9	74.5	61.9	86.5
ReCap-DataComp-1B v2	56.5	13.1	303	1451/310	67.8	74.7	61.2	86.6

Keep Ratio	Text VQA	Chart QA	OCR.	MME	SEED	SQA	GQA	POPE
100%	53.8	12.2	254	1409/350	65.9	73.9	60.3	84.7
90%	56.3	12.4	266	1461/335	67.6	74.8	61.1	85.4
75%	55.8	13.1	293	1438/283	68.6	73.9	61.7	86.3
50%	55.4	12.8	299	1429/313	68.5	73.8	61.6	86.5
35%	56.9	12.9	291	1426/293	67.9	74.5	61.9	86.5
25%	56.7	12.5	283	1430/297	67.8	76.3	61.4	86.3
10%	55.6	13.0	276	1412/301	66.1	75.0	61.2	85.4

A higher keep ratio retains more vision tokens as captioning conditions, while a lower keep ratio masks more tokens

Discussion

Loss of Contrastive Signal / Alignment Robustness

- Dropping contrastive image-text and relying only on generative captions weakens alignment robustness, hurting retrieval, zero-shot discrimination, and fine-grained image-text matching when captions are noisy and biased.

Reliance on Synthetic Captions

- Heavy reliance on synthetic captions bakes in their quality, biases, omissions, and style; because they focus on silent, generic content rather than exhaustive scene detail, coverage for rare objects, fine-grained attributes, and complex relationships can suffer

Caption-Only Objective Might Miss Non-Descriptive Visual Features

- A caption-only objective neglects non-descriptive visual cues (e.g., low-level textures, subtle spatial relations, background details) that contrastive learning can capture, potentially reducing performance on tasks needing these features.