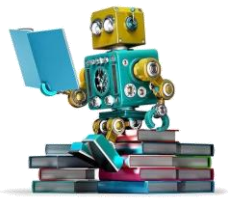
SDML
ML Paper Review

May 2025



Describe Anything: Detailed Localized Image and Video Captioning

Long Lian et al., Nvidia+

https://arxiv.org/abs/2504.16072

NVIDIA

Describe Anything: Detailed Localized Image and Video Captioning

Long Lian[1,2] Yifan Ding[1] Yunhao Ge[1] Sifei Liu[1] Hanzi Mao[1] Boyi Li[1,2] Marco Pavone[1]
Ming-Yu Liu[1] Trevor Darrell[2] Adam Yala[2,3] Yin Cui[1]
[1]NVIDIA [2]UC Berkeley [3]UCSF

Figure 1: **Describe Anything Model (DAM)** generates **detailed localized captions** for user-specified regions within **images** (top) and **videos** (bottom). DAM accepts various region specifications, including clicks, scribbles, boxes, and masks. For videos, specifying the region in *any frame* suffices.

Abstract

Generating detailed and accurate descriptions for specific regions in images and videos remains a fundamental challenge for vision-language models. We introduce the Describe Anything Model (DAM), a model designed for detailed localized captioning (DLC). DAM preserves both local details and global context through two key innovations: a focal prompt, which ensures high-resolution encoding of targeted regions, and a localized vision backbone, which integrates precise localization with its broader context. To tackle the scarcity of high-quality DLC data, we propose a Semi-supervised learning (SSL)-based Data Pipeline (DLC-SDP). DLC-SDP starts with existing segmentation datasets and expands to unlabeled web images using SSL. We introduce DLC-Bench, a benchmark designed to evaluate DLC without relying on reference captions. DAM sets new state-of-the-art on 7 benchmarks spanning keyword-level, phrase-level, and detailed multi-sentence localized image and video captioning.

1. Introduction

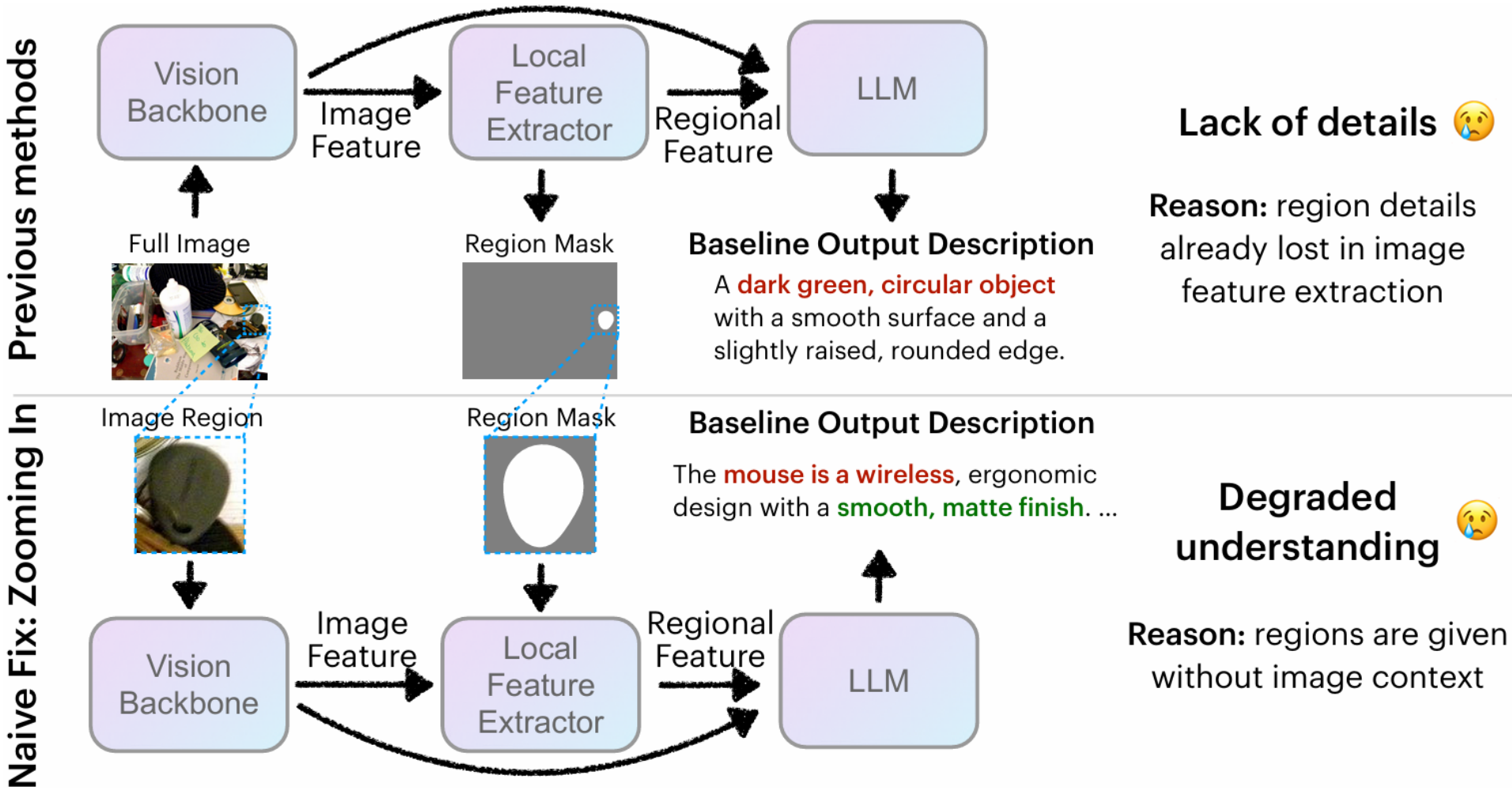
Image captioning has been a longstanding challenge in computer vision and natural language processing [18], as it involves understanding and describing visual content in natural language. While recent Vision-Language Models (VLMs) have achieved impressive results in image-level captioning, generating detailed and accurate captions for specific regions within an image remains an open problem. This challenge intensifies with videos, where models must additionally capture dynamic visual content, such as human actions, object motions, and human-object interactions. If resolved, it would open new doors for fine-grained grounded image/video understanding [49, 102] and generation [40, 42].

Most existing VLMs (*e.g.*, GPT-4o [54]) lack mechanisms for precise localization. Recent approaches that empower VLMs to take 2D localization cues such as bounding boxes [30, 82, 85, 101] often yield brief phrases rather than detailed descriptions. While there are methods [15, 45, 93, 95, 99, 100] that produce longer captions, they provide minimal detail or in-

Describe Anything overview

- Describe Anything Model (DAM) solves the problem of detailed localized captioning
  - Describing individual objects in an image or video, not the whole thing
  - Ability to provide rich object details without including extraneous details
- Prior work with full images tended to be short or vague and/or include other image/video info that is not relevant
- Prior work with crops tended to lose contextual cues, hurting accuracy, sometimes to the point of complete recognition failure
- DAM built a novel input backbone which fed into a pre-trained LLM
  - Clever design avoided needing to train the multimodal LLM from scratch

Prior methods



Examples

Figure 1: **Describe Anything Model (DAM)** generates **detailed localized captions** for user-specified regions within **images** (top) and **videos** (bottom). DAM accepts various region specifications, including clicks, scribbles, boxes, and masks. For videos, specifying the region in *any frame* suffices.

Classic deep learning dilemma

Detailed localized captioning faced the typical trio of problems:

- Need a good architecture
  - As mentioned, neither whole images/videos nor crops worked well
- Need training data
  - Datasets only had short phrases
  - Additional problem that bounding boxes that identified the object of interest were sometimes ambiguous
- Need to measure performance accurately
  - Short reference captions unfairly penalized extra correct details
  - LLMs didn’t seem to have enough understanding for LLM-based scoring

Describe Anything solutions

To solve the trio of problems, addressed each of them in the paper:

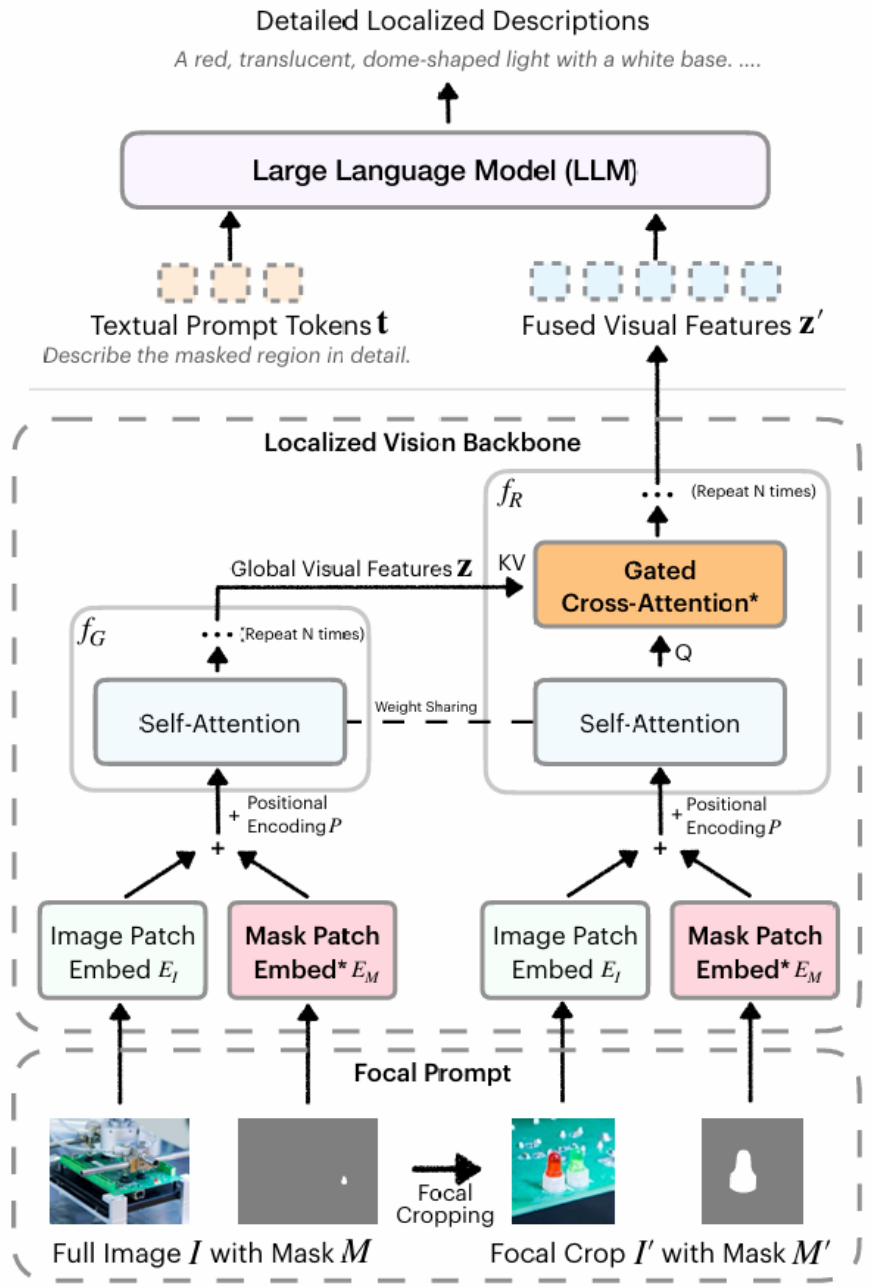
- Describe Anything Model (DAM) architecture
  - Combines full image/frame and crops – along with segmentation masks
  - Their localized vision backbone combines all of the above and feeds the LLM
- Semi-supervised learning (SSL)-based Data Pipeline (DLC-SDP)
  - Included self-training with unlabeled web images
- DLC-Bench
  - Uses LLM to score captions based on positive and negative criteria

Model inputs

- Formulated task with inputs as image and binary mask pairs
  - If mask not available, existing segmentation models like SAM(2) can turn points, boxes, or scribbles into segmentation masks
  - For videos, use a sequence of *N* of these pairs
  - Existing models can also track an object from a single video frame onto other frames, s.t. indicating the region of interest once is sufficient
- DAM takes in full image and full mask
- It also takes in a crop 3x the object size from image and from mask
  - For very small objects, a min size was 48 pixels in each dimension

Model architecture

- DAM created a *localized vision backbone*
- Used SigLIP pre-trained ViT encoder on the combined full image and mask
- Used a modified SigLIP on the combined crop and mask
  - An extra cross-attention layer brings global context information from the encoded full image and mask
- Masks and cross-attention have scalar gating that starts at zero, so initially during training only crop image is used



Model training

- Builds on the vision-language training from the VILA paper
  - Uses VILA 1.5 available on Hugging Face (based on LLaVA)
  - Fine-tuned VILA 1.5 3B for images, and VILA 1.5 8B for videos
- Vison backbone outputs same number of tokens as original SigLIP
  - Despite having four inputs, doesn't increase sequence length of the vision tokens
- End to end training of vision backbone, token projector, and LLM
- Mixed ShareGPT-4V text task with the image/video captioning task, to avoid catastrophic forgetting of text capabilities
- Starting with pre-trained models, training time was moderate
  - The 3B image model was trained on 8 A100 GPUs in under one day
  - The 8B video model was trained on 32 A100 GPUs in under one day

Prompt augmentation

- Wanted ability to produce shorter and longer captions
- 20% of prompts had suffix “in X sentence(s)” based on target caption
- 20% of prompts had suffix using word counts
  - For short targets, used exact counts, such as “in 3 words”
  - For medium rounded to nearest ten, such as “in about 50 words”
  - For very long, used “in more than 200 words”

1. Describe the masked region {prompt_suffix}.
2. Describe the masked area {prompt_suffix}.
3. What can you describe about the masked region {prompt_suffix}?
4. Can you describe the masked region {prompt_suffix}?
5. Provide an explanation of the masked region {prompt_suffix}.
6. Depict the masked area {prompt_suffix}.
7. Portray the masked area {prompt_suffix}.
8. Describe what the masked region looks like {prompt_suffix}.
9. Illustrate the masked region {prompt_suffix}.
10. How would you explain the masked area {prompt_suffix}?
11. What details can you provide about the masked region {prompt_suffix}?
12. What does the masked region entail {prompt_suffix}?
13. How would you illustrate the masked region {prompt_suffix}?
14. How would you depict the masked area {prompt_suffix}?
15. How would you portray the masked area {prompt_suffix}?

SSL-based data pipeline

- Semi-supervised learning (SSL)-based Data Pipeline (DLC-SDP)
- Stage 1 takes existing segmentation datasets, which have short object classes or other name, and finds VLMs already can accurately expand these to longer, detailed descriptions. Training starts with this data.
  - DAM is trained to predict the detailed captions without the classes/labels
- Stage 2 takes the early version of DAM and iterates:
  - Open-vocabulary segmentation models generate mask from web images
  - DAM generates pseudo-labeled predictions
  - Confidence-based filtering keeps only high-quality (high similarity) predictions
  - The kept triplets (image, mask, description) are added to training set

DLC-Bench [1]

- New benchmark eliminated need for comprehensive reference captions
- Curated positive and negative questions about the region
- Used LLM as a judge to compare whether the predictions matched positives (gaining points) and/or matched negatives (losing points)
- Model should not be penalized for how it words the details
- Model will not be falsely penalized for including accurate details that weren’t included in a reference caption
- If model includes incorrect or extraneous information, it will, on average, trigger some of the negative questions and lose points

DLC-Bench [2]

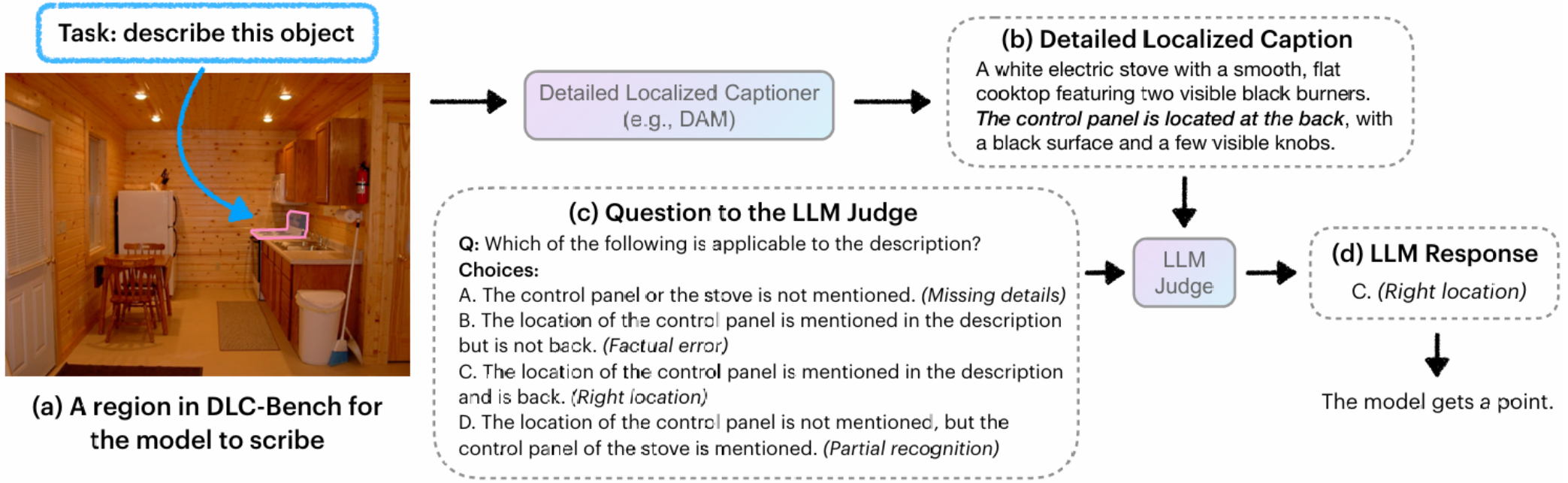


Figure 4: **We propose DLC-Bench, a benchmark tailored to detailed localized captioning.** In DLC-Bench, a captioning model is prompted to describe a specified image region (a). The generated description (b) is then evaluated by querying an LLM Judge (c). Points are assigned or deducted based on the LLM's response (d). The question we show in (c) is an example of positive questions.

Results [1]

- DAM achieves SOTA and does significantly better on existing benchmarks, despite the shortcomings of measuring keyword, phrases, or reference captions

Method	LVIS (%)		PACO (%)	
	Sem. Sim. (↑)	Sem. IoU (↑)	Sem. Sim. (↑)	Sem. IoU (↑)
LLaVA-7B [48]	49.0	19.8	42.2	14.6
Shikra-7B [15]	49.7	19.8	43.6	11.4
GPT4RoI-7B [99]	51.3	12.0	48.0	12.1
Osprey-7B [95]	65.2	38.2	73.1	52.7
Ferret-13B [93]	65.0	37.8	-	-
VP-SPHINX-7B [45]	86.0	61.2	74.2	49.9
VP-LLAVA-8B [45]	86.7	61.5	75.7	50.0
DAM-8B (Ours)	**89.0**	**77.7**	**84.2**	**73.2**

Table 2: LVIS [29] and PACO [60] open-class **keyword-level** captioning benchmarks. DAM excels particularly in the challenging PACO benchmark that requires distinguishing between objects and parts.

Method	BLEU	METEOR	ROUGE-L	CIDEr	SPICE
Shikra-7B [15]	18.2	15.3	25.2	49.8	22.0
GPT4RoI-7B [99]	19.7	17.7	29.9	61.7	24.0
Ferret-7B [93]	11.1	8.8	22.7	38.1	17.5
VP-SPHINX-13B [45]	15.2	15.6	27.2	67.4	24.0
RegionGPT-7B [28]	16.1	16.7	27.4	54.6	20.5
DAM-8B (Ours)	**22.6**	**17.8**	**31.2**	**74.7**	**25.5**

Table 3: Zero-shot evaluation on **phrase-level** dataset Flickr30k Entities [58]. Our model achieves 12.3% average relative improvement against previous best.

	Short Captioning Metrics					Long Cap. Metrics
Method	BLEU	METEOR	ROUGE-L	CIDEr	SPICE	CLAIR
Shikra-7B [15]	29.5	11.1	23.9	42.7	9.0	34.5
GPT4RoI-7B [99]	27.1	11.6	26.8	59.9	11.1	43.9
Ferret-7B [93]	24.6	10.7	22.3	39.7	8.2	45.2
GLaMM-7B [61]	23.2	10.1	23.8	51.1	8.7	43.8
VP-SPHINX-13B [45]	22.6	10.7	22.6	32.4	7.6	51.2
RegionGPT-7B [28]	25.4	12.2	25.3	42.0	8.1	37.2
DAM-8B (Ours)	**38.7**	**19.4**	**37.1**	**70.0**	**16.9**	**57.9**

Table 4: Zero-shot evaluation on the **detailed captioning** dataset Ref-L4 [14]. Our method achieves 33.4% and 13.1% average relative improvement on the short/long language-based captioning metrics, respectively.

Results [2]

- The new DLC-Bench shows SOTA, including beating GPT-4o and o1 (right)
- DAM also achieves top scores for video captioning on the HC-STVG benchmark (below)

Method	BLEU@4	METEOR	ROUGE-L	CIDEr	SPICE
Osprey-7B [95]	0.7	12.0	18.0	1.2	15.6
Ferret-13B [93]	0.5	10.2	17.0	1.2	11.2
Shikra-7B [15]	1.3	11.5	19.3	3.1	13.6
Merlin-7B [94]	3.3	11.3	26.0	10.5	20.1
Artemis-7B [59]	15.5	18.0	40.8	53.2	25.4
VideoRefer-7B [96]	16.5	18.7	42.4	68.6	28.3
DAM-8B (Ours)	**19.8**	**21.0**	**45.9**	**91.0**	**31.4**

Method	#Params	Pos (%)	Neg (%)	Avg (%)
General VLMs:				
GPT-4o [54]	-	43.4	79.6	61.5
o1 [55]†	-	46.3	78.8	62.5
Claude 3.7 Sonnet [73]†	-	21.8	50.4	36.1
Gemini 2.5 Pro [74, 75]†	-	36.5	75.2	55.8
Llama-3.2 Vision [25]	11B	30.7	63.8	47.3
VILA1.5-Llama-3 [44]	8B	22.5	61.0	41.8
InternVL2.5 [20, 21, 84]	8B	15.9	42.0	28.9
LLaVA v1.6 [46–48]	7B	15.4	55.0	35.2
Qwen2.5-VL [77, 81]	7B	20.3	62.2	41.2
VILA1.5 [44]	3B	16.0	50.0	33.0
Region-specific VLMs (full / cropped input):				
GPT4RoI [99]	7B	6.5/3.5	46.2/52.0	26.3/27.7
Shikra [15]	7B	2.7/8.0	41.8/51.4	22.2/29.7
Ferret [93]	7B	6.4/14.2	38.4/46.8	22.4/30.5
RegionGPT [28]	7B	13.0/10.6	41.4/46.4	27.2/28.5
ControlCap [101]	0.3B	18.3/ 3.6	75.6/53.6	47.0/28.6
SCA [30]	3B	3.4/ 0.1	44.6/18.4	24.0/ 9.3
OMG-LLaVA [100]	7B	0.9/ 5.6	16.0/32.6	8.5/19.1
VP-SPHINX [45]	13B	11.7/26.3	33.2/71.6	22.5/49.0
DAM (Ours)	3B	**52.3**	**82.2**	**67.3**

Describe Anything conclusion

- DAM achieves SOTA on 7 benchmarks for localized captioning
- Had to solve the common trio of problems by:
  - Designing the localized vision backbone architecture
  - Iterating on a semi-supervised learning process to generate a large training dataset
  - Creating a new DLC-Bench benchmark to accurately measure performance
- Design choices allowed them to use a pre-trained image encoder and pre-trained multimodal LLM without needing much retraining
  - This also reduced the amount of training data needed
- Training time was reasonable, under 1 day on 8 or 32 A100s, for the 3B image and the 8B video models, respectively
- See more details in the project page, paper, or code repo https://describe-anything.github.io/

References

- VILA: On Pre-training for Visual Language Models
  Ji Lin, et al. (2023)
  https://arxiv.org/abs/2312.07533
- Sigmoid Loss for Language Image Pre-Training
  Xiaohua Zhai, et al. (2023)
  https://arxiv.org/abs/2303.15343