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-

clude unrelated content from other regions, as shown in Fig. 5. This raises the question: *What makes detailed localized captioning (DLC) so challenging?*

We identify three key obstacles to DLC:

1. **Loss of Region Details:** As shown in Fig. 2, prior methods extract local features from global image representations, often leading to loss of fine-grained details—particularly for small objects in complex scenes. By the time the LLM processes the visual features, crucial details necessary for generating precise captions are already lost. Cropping the region of interest may enhance detail but risks losing essential contextual cues.
2. **Scarcity of High-Quality Datasets:** Datasets such as RefCOCOs [31, 50] and Visual Genome [34] typically offer only short phrases that do not suffice for training models to generate rich, detailed captions. Recent synthetic data approaches [45, 93] are based on bounding boxes that could not precisely convey the exact region of interests, while methods that rely on global captions [28] may have difficulty capturing non-salient regions.
3. **Limitations in Benchmarks:** Prior localized captioning benchmarks compare generated captions against reference captions using language-based image captioning metrics [2, 5, 43, 56, 79] or LLM-based scoring [95, 96]. However, such techniques are not very applicable to DLC. Since the reference captions provided in the benchmarks often lack comprehensive details of the region, DLC models are often unfairly penalized for correct details not explicitly mentioned in the reference.

We propose the following solutions to these challenges:

To tackle the loss of details in regional features, we propose the **Describe Anything Model** (DAM), which preserves both local detail and global context. DAM achieves this through two key innovations (Fig. 3): **1)** the *focal prompt*, which encodes the region of interest with high token density, and **2)** the *localized vision backbone*, which ensures precise localization while integrating global context. These components enable DAM to generate detailed and accurate captions, even for small objects in complex scenes.

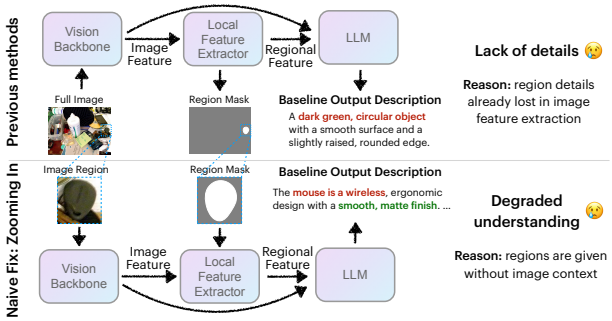


Figure 2: **Top:** Prior regional captioners derive regional features from global image representations, leading to vague descriptions. **Bottom:** Zooming in (cropping the image region) enhances detail but loses contextual cues, degrading recognition. This underscores the need for a design that **encodes detail-rich regional features while preserving context for improved DLC performance**.

To overcome the scarcity of high-quality DLC datasets, we introduce **Semi-supervised learning (SSL)-based Data Pipeline** (DLC-SDP) to generate high-quality localized captions in two stages. First, leveraging high-quality masks and keywords (*e.g.*, class names, part names, or entities) from human-annotated segmentation datasets, we query a VLM to expand each keyword into a detailed caption given each mask-referred region. Second, inspired by self-training-based SSL in image classification [7, 8, 37, 69, 88], DLC-SDP performs self-training with web images as an unlabeled dataset and segmentation datasets as labeled data. An LLM further summarizes descriptions into multiple granularities, yielding a diverse dataset with high-quality localized captions, enabling our model to outperform strong API-only baselines such as GPT-4o [54] and o1 [55].

To mitigate the limitations of current benchmarks, we introduce **DLC-Bench**, which evaluates detailed localized captions based on a set of predefined positive and negative attributes for each region, eliminating the reliance on comprehensive reference captions. This approach provides a more flexible and accurate evaluation, encouraging models to generate informative and precise descriptions.

We summarize our contributions as follows:

1. **Describe Anything Model (DAM):** A novel architecture with a focal prompt and a localized vision backbone for multi-granular regional image and video captioning.

Component	Previous Practice	Problem	Our Solution	Advantages
Describe Anything Model (DAM)	Extracting regional features from global image features	Regional details already lost in image feature extraction and not provided to the LLM	Providing **focal prompt** to proposed **localized vision backbone**	Detail-rich contextful features allowing for accurate, multi-granular localized descriptions
SSL Data Pipeline (DLC-SDP)	Query a data curation VLM with referring **boxes** and **global image captions**	Imprecise referring to data curation model	Reframe the query into a **mask**-referred **keyword** expansion question	Leverage high-quality precise human annotated regional masks and keywords
	Fully supervised learning	Limited data with high annotation quality	**Semi-supervised learning**	Scalable to diverse web-scale unlabeled datasets
Benchmark (DLC-Bench)	Pred caption + **reference GT caption** → language-based similarity metrics or LLM scorer	Incorrect hallucination penalty for correct details not present in the reference caption	Pred caption + **query for positive/negative attributes** → LLM scorer	Accurate detail and hallucination assessment without relying on reference captions

Table 1: Advantages of our proposed model DAM, our SSL data pipeline DLC-SDP, and our benchmark DLC-Bench to previous practices.

2. **SSL Data Pipeline (DLC-SDP):** A semi-supervised data pipeline that leverages high-quality segmentation annotations and unlabeled web images for scalable and diverse data curation.
3. **DLC-Bench:** A benchmark designed to evaluate DLC without reference captions.

Unlike generalist models, we focus on *localized image and video captioning* across *multiple granularities*, **achieving SOTA performance on 7 benchmarks** across keyword, phrase, and detailed multi-sentence captioning. Advantages of DAM, DLC-SDP, and DLC-Bench to prior practices are presented in Tab. 1. We release our code, models, data, and benchmark at describe-anything.github.io.

2. Related Work

Vision-Language Models (VLMs). VLMs integrate visual and textual inputs for multimodal understanding and are broadly classified into BLIP-style [1, 3, 22, 25, 36, 38, 39] and LLaVA-style [4, 6, 9, 19, 20, 44, 46–48, 81, 87, 103]. However, these models lack precise localization capabilities, limiting their ability to generate regional descriptions.

Localized Image Captioning. While general VLMs generate image-level captions, localized captioning requires fine-grained regional descriptions. SoM [91, 92] augments VLMs with visual markers, but these markers may blend with the background, as discussed in App. A. Region-aware VLMs [15, 28, 30, 45, 57, 61, 70, 82, 83, 95, 99–101] introduce regional referring inputs. Recent efforts such as Merlin, Artemis, and VideoRefer [59, 94, 96] extend region-based captioning to videos. However, these methods still struggle with capturing intricate details in the referring regions, as shown by the examples in Fig. 5. This is because prior models either extract localized features from global image embeddings or simply encode the referring condition as referring tokens, which leads to insufficient regional details for the LLM, especially for small objects. We address this via focal prompting and a localized vision backbone, balancing local detail with global context.

Another limitation is the scarcity of high-quality datasets. Datasets like RefCOCOs [31, 50] and VG [34] provide only short phrases. Recent approaches [28, 45, 61, 93, 95, 96] use bounding-box-based VLM queries, sometimes augmented with global captions, for synthetic generation, which leads to caveats discussed in App. C.1. We propose an SSL data pipeline that uses human-annotated and unlabeled data for richer regional descriptions.

Benchmarking Localized Captioning. [28, 30, 59, 61, 93, 95, 98, 100, 101] evaluate localized captioning by computing language-based image captioning metrics [2, 5, 43, 56, 79] between predicted captions and reference captions. However, these metrics focus on textual matching and may not correlate well with the factual correctness or quality for detailed descriptions. [95, 96] use Sentence-BERT [63] and text-only LLMs to score the predictions against reference captions. However, reference captions often lack comprehensive details about the region of interest, which penalizes models for correct details not explicitly mentioned in the reference by treating them as hallucinations, as discussed in App. C.2. Our DLC-Bench resolves this issue by eliminating the need for reference captions.

Vision Models with Focus. Prior works enhance attention to salient regions using focal self-attention [90], ToSA [68], ToMe [10], DW-ViT [65], Quadformer [66], and V∗ [86]. These methods allocate resources dynamically to salient regions defined by the model. In contrast, our focal prompt explicitly prioritizes user-specified regions, ensuring accurate and detailed captions even for non-salient objects.

3. DAM: Describe Anything Model

Describe Anything Model (DAM) generates detailed localized descriptions of user-specified regions within images and videos. DAM effectively balances local detail and contextual information through our proposed *focal prompt* and *localized vision backbone*.

3.1. Task Formulation

The task of detailed localized captioning involves generating comprehensive textual descriptions focused exclusively on specified regions within images or videos. Formally, given N input frames $I^{(i)} \in \mathbb{R}^{H\times W\times 3}$ and corresponding binary masks $M^{(i)} \in \{0,1\}^{H\times W}$ indicating the region of interest in each frame, the objective is to produce a detailed description T of the content within the region through a captioning model:

$$T = \text{CaptioningModel}\left(\{I^{(i)}, M^{(i)}\}_{i=1}^{N}\right) \quad (1)$$

We focus on using binary masks $M^{(i)}$ as the localization input, since other forms of localization (*e.g.*, points, scribbles, boxes, or masks on an image or a subset of frames in a video) can be transformed into masks via segmentation models such as SAM [32] and SAM 2 [62]. For simplicity, we first introduce our method for localized image captioning, omitting the frame index i, and later extend it to videos in Sec. 3.3.

3.2. Model Architecture

As shown in Fig. 3, DAM consists of two key components: the *focal prompt* and the *localized vision backbone*.

3.2.1. Focal Prompt

To provide a detailed representation of the region of interest within its context, we introduce the *focal prompt*, which includes both the full image and a focal crop centered around the specified area, along with their corresponding masks.

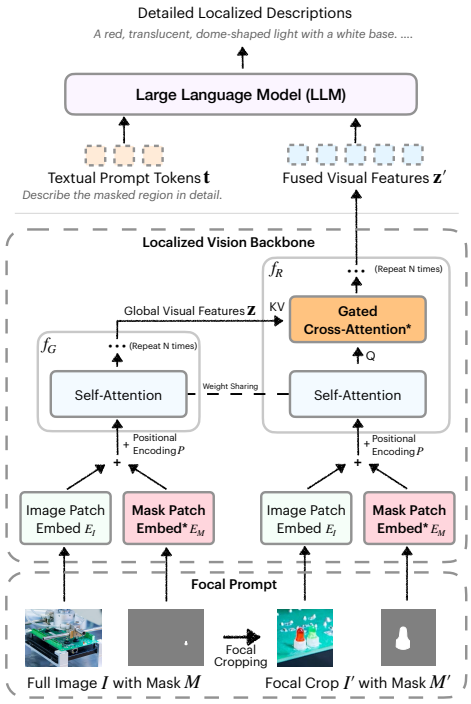


Figure 3: **Architecture of the Describe Anything Model (DAM).** DAM employs a *focal prompt* to encode user-specified regions with high token density while preserving context for detailed understanding. Focal cropping is applied to the image and its corresponding mask, retaining surrounding areas for local context. Both the full image and the focal crop are the inputs into the *localized vision backbone*, where images and binary masks are embedded in a spatially aligned fashion. Global context from the full image is leveraged to help understand the focal crop through gated cross-attention. The resulting visual features and prompt tokens are fed into a large language model to generate detailed, context-aware descriptions. * indicates initialized to output zeros.

We first extract the bounding box B of the mask M and expand it by a factor α in both the horizontal and vertical directions to include additional surrounding context:

$$B' = \text{ExpandBox}(B, \alpha). \quad (2)$$

For instance, setting $\alpha = 3$ results in a region that can be up to 9 times as large as the original bounding box, subject to clipping at the image boundaries. If either the height or width of the expanded box is less than 48 pixels, we enforce a minimum size of 48 pixels in that dimension to ensure sufficient context for very

small regions.

The focal crop of the image and mask are then:

$$I' = I|B', \quad M' = M|B', \tag{3}$$

where $|B'$ denotes cropping to B'. The focal prompt thus consists of 1) the full image I and its mask M and 2) the focal crop I' and its mask M'. By including both the full image and the focal crop, along with their masks, the focal prompt contains both global context and a detailed view of the region of interest.

3.2.2. Localized Vision Backbone

Effectively processing all four components of the focal prompt with a VLM is non-trivial, as naively concatenating the full image and the focal crop leads to a loss in performance (Tab. 8). We propose the *localized vision backbone*, which 1) achieves localized understanding by encoding the masks in a spatially aligned manner and 2) integrates global context into the region of interest through gated cross-attention.

Handling Localization Inputs. Similar to how an image is encoded by a linear patch embedding layer in vision transformers (ViTs) [24], we integrate the mask M into its corresponding full image I through another patch embedding layer that takes in 2D inputs with one channel.

Specifically, the full image I and its mask M are processed through patch embedding layers, followed by the global vision encoder $f_{\text{G}}(\cdot)$ to obtain global visual features $\mathbf{z}$. The focal crop I' and its mask M' undergo a similar process with the regional vision encoder $f_{\text{R}}(\cdot)$, except that $f_{\text{R}}(\cdot)$ also takes $\mathbf{z}$ as a context to obtain the final fused visual features $\mathbf{z}'$. Specifically, we have:

$$\mathbf{x} = E_{\text{I}}(I) + E_{\text{M}}(M) + P, \quad \mathbf{z} = f_{\text{G}}(\mathbf{x}), \tag{4}$$

$$\mathbf{x}' = E_{\text{I}}(I') + E_{\text{M}}(M') + P, \quad \mathbf{z}' = f_{\text{R}}(\mathbf{x}', \mathbf{z}), \tag{5}$$

where $E_{\text{I}}(\cdot)$ and $E_{\text{M}}(\cdot)$ are the image and mask patch embedding layer, respectively, $\mathbf{x}$ and $\mathbf{x}'$ are global and focal embedded inputs with information for both the image and the mask, and P denotes the positional encoding.

The newly added mask embedding layer E_{M} is initialized to output zeros, ensuring that the VLM's initial behavior is unaffected prior to fine-tuning.

Regional Feature Encoding with Gated Cross-Attention Adapters. To integrate global context into the focal prompt, we insert gated cross-attention adapters [1, 38] into each transformer block of the regional vision encoder f_{R}. After the self-attention and feed-forward layers, we add a gated cross-attention mechanism that allows local features to attend to global features:

$$\mathbf{h}^{(l)'} = \mathbf{h}^{(l)} + \tanh\left(\gamma^{(l)}\right) \cdot \text{CrossAttn}\left(\mathbf{h}^{(l)}, \mathbf{z}\right), \tag{6}$$

$$\mathbf{h}^{(l)}_{\text{Adapter}} = \mathbf{h}^{(l)'} + \tanh\left(\beta^{(l)}\right) \cdot \text{FFN}\left(\mathbf{h}^{(l)'}\right), \tag{7}$$

where $\mathbf{h}^{(l)}$ is the output of the l-th self-attention block in f_{R}, $\gamma^{(l)}$ and $\beta^{(l)}$ are learnable scaling parameters initialized to zero, and CrossAttn denotes cross-attention with queries from $\mathbf{h}^{(l)}$ and keys and values from the global features $\mathbf{z}$, similar to how cross-attention is employed in encoder-decoder Transformers [78]. $\mathbf{h}^{(l)}_{\text{Adapter}}$ is used in place of $\mathbf{h}^{(l)}$ in the next Transformer block. To reduce the number of parameters, f_{R} shares self-attention block weights with f_{G}.

By initializing $\gamma^{(l)}$ and $\beta^{(l)}$ to zero, we ensure that the initial behavior of the model remains identical to the original VLM prior to fine-tuning. During training, the model learns to leverage the global context to enhance local feature representations, facilitating detailed and contextually accurate descriptions.

Generating Detailed Localized Descriptions. The visual features from both the global and regional vision encoders are combined and fed into the large language model to generate detailed, context-aware descriptions T:

$$T = \text{LLM}(\mathbf{t}, \mathbf{z}'), \tag{8}$$

where $\mathbf{t}$ denotes textual prompt tokens.

Notably, the proposed components do *not* increase the sequence length of the vision tokens, ensuring that DAM remains efficient. By initializing new modules (mask embedding E_{M} and scaling parameters $\gamma^{(l)}$ and $\beta^{(l)}$) to zeros, we preserve the pre-trained capabilities of the VLM prior to fine-tuning, *allowing for smooth adaptation of an off-the-shelf VLM without rerunning pre-training*. Thanks to this design, our model requires way less training data (~ 1.5M samples) than prior works that involve VLM pretraining.

3.3. Extension to Videos

Since images can be considered as videos with a single frame, the model naturally extends to handling videos by processing sequences of frames and their corresponding masks. The visual features from all frames are concatenated in the sequence dimension and fed into the language model to generate detailed localized descriptions across the video frames, compatible with how VLMs are pretrained to handle videos. We leverage SAM 2 [62] to turn sparse localizations into a mask for each frame.

4. DLC-SDP: SSL-based Data Pipeline

The effectiveness of DAM depends critically on the availability of high-quality training data for detailed localized descriptions. To this end, we propose the **Semi-supervised learning (SSL)-based Data Pipeline** (DLC-SDP), a two-stage approach that enables us to build a large and diverse dataset with high-quality localized descriptions.

4.1. Stage 1: Leveraging Existing Annotations

The first stage of DLC-SDP reframes the data generation problem into a *vision-grounded description expansion task*. We observed that although current VLMs often struggle to generate detailed localized descriptions when given a referring mask, they can effectively expand short localized descriptions into detailed ones. Inspired by this observation, we leverage high-quality human-annotated masks and keywords (object class names, part names, entities, etc.) from existing segmentation datasets, reframing the VLM query to expand each regional keyword into a detailed caption given the referring masks.

Importantly, our model is trained to predict these high-quality descriptions *without* taking the initial keywords as inputs. Since there are no class labels provided at inference time–and existing VLMs perform poorly without them–our approach ensures superior data quality compared to direct VLM prompting for distillation.

4.2. Stage 2: SSL with Unlabeled Data

Since it is not scalable to rely on high-quality manual annotations, the second stage of DLC-SDP employs self-training-based semi-supervised learning techniques inspired by successful approaches in image classification [7, 8, 69, 89]. Our self-training approach involves four steps:

1. **Mask Generation.** We use open-vocabulary segmentation models [30, 44] to extract object masks from unlabaled web images.
2. **Description Generation.** Our DAM, initially trained on the DLC dataset based on annotated segmentation datasets, generates detailed localized descriptions for these regions.
3. **Confidence-based Filtering.** We apply CLIP-based confidence filtering to keep only high-quality samples, following SSL literature.
4. **Data Expansion.** The newly generated (image, mask, description) triplets are added to our training dataset.

This semi-supervised approach dramatically expands the range of object categories and increases data diversity beyond the initial supervised dataset. Furthermore, to support the capability of DAM for multi-granular captioning, we leverage an LLM to summarize the detailed descriptions into shorter forms (*e.g.*, phrases or short sentences), enabling DAM to flexibly generate captions ranging from succinct phrases to multi-sentence narratives.

By systematically curating our training data through this two-stage SSL pipeline, DAM achieves significant performance improvements. Notably, our model trained on data obtained with DLC-SDP outperforms GPT-4o [54] and o1 [55], two strong closed-sourced baselines, showing the effectiveness of DLC-SDP. We present implementation details of DLC-SDP in App. H.

5. DLC-Bench: Benchmark for DLC

We introduce **DLC-Bench**, a benchmark designed for DLC to eliminate the need for comprehensive reference captions. The core intuition behind DLC-Bench is that an ideal description should be rich in relevant details while strictly avoiding factual errors or information for irrelevant regions. Therefore, we assess predictions based on a set of predefined positive and negative attributes for each region.

As illustrated in Fig. 4, the evaluation process for a model like DAM has two steps:

1. The model is prompted to generate a detailed description for each masked region in the benchmark

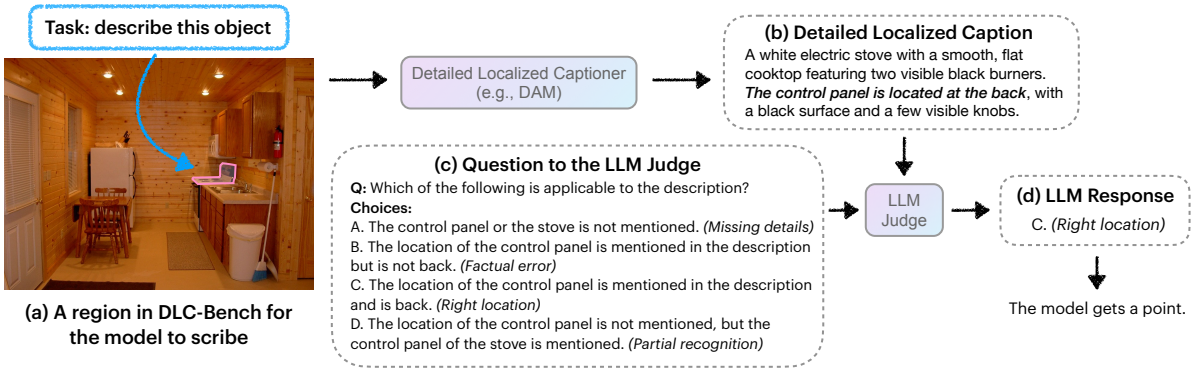


Figure 4: **We propose DLC-Bench, a benchmark tailored to detailed localized captioning.** In DLC-Bench, a captioning model is prompted to describe a specified image region (a). The generated description (b) is then evaluated by querying an LLM Judge (c). Points are assigned or deducted based on the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

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.

Method	Short Captioning Metrics					Long Cap. Metrics
	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.

dataset.

2. An LLM serves as a judge, assessing the generated description by responding to a set of manually curated positive and negative questions about region details.

DLC-Bench employs two categories of questions for each annotated instance, with example in Fig. A.5:

- **Positive questions** focus on specific attributes of object parts that *should* be present in the description. The model earns a point if the description accurately includes the specified detail; omissions receive no points, while factual errors incur a penalty.
- **Negative questions** focus on details that *should not* be present–either attributes typical of similar objects but absent in the target instance, or descriptions irrelevant to the specified region. A point is awarded if the model correctly omits such details; conversely, including them results in a penalty. To avoid getting high scores for captions that are completely off, point could only be awarded if the caption has the correct recognition of the object.

This approach provides a more flexible and accurate evaluation, encouraging models to generate informative and precise descriptions without constraints from incomplete reference captions.

Our DLC-Bench comprises a total of 892 manually verified questions covering a wide range of attributes and potential cases for hallucinations. Details on the curation process and the scoring mechanism are provided in App. D.

6. Results

DAM excels at *localized image and video captioning* across *multiple granularities* including keyword, phrase, and detailed captions, **achieving SOTA on 7 in-domain and zero-shot benchmarks** (Tabs. 2 to 7). We explain the details for each benchmark in App. B.

6.1. Quantitative Results

Open-class keyword-level localized captioning task requires the model to output keywords containing the object and part entities to describe the region. Tested on object-level LVIS [29] and part-level PACO [60]

datasets in Tab. 2, our method achieves state-of-the-art performance. In the PACO benchmark, a challenging benchmark that includes both full objects and parts in complex scenes and requires the model to decide whether the region is an object or a part, our method achieves 73.2% semantic IoU and 84.2% semantic similarity, outperforming the previous best by 23.2% and 8.5% respectively.

Phrase-level localized captioning task requires the model to output a phrase containing a brief description for each region that includes object identification and attributes typically within a few words. Tested zero-shot on Flickr30k Entities [58], our model achieves strong performance, outperforming the previous best by 12.3% relative improvement on an average of 5 metrics.

Detailed localized captioning task requires the model to output a detailed description for each the region with the length spanning from a long sentence to multiple sentences. An ideal description includes the description of the object in the region, its parts, as well as their attributes and relationships. We benchmark this capability on the challenging Ref-L4 [14] benchmark and our proposed DLC-Bench. On the Ref-L4 benchmark in Tab. 4, our method achieves 33.4% relative improvement on average over the previous best on short language-based captioning metrics [2, 5, 43, 56, 79], and 13.1% relative improvement on the long language-based captioning metrics [13].

We also benchmark various regional captioning models on **our proposed DLC-Bench**, which does not suffer from the limitations of requiring reference captions in previous benchmarks. As shown in Tab. 5, our Describe Anything Model (DAM) significantly outperforms existing general and region-specific VLMs, achieving state-of-the-art positive and negative accuracy and demonstrating its ability to produce detailed and accurate descriptions. Remarkably, DAM surpasses GPT-4o [54] and o1 [55], two strong API-only baselines. DAM also surpasses models with thinking mode enabled [55, 73, 74].

Detailed localized video captioning requires the model to output a detailed description for each region in a video. We benchmark this capability on the challenging HC-STVG [71] benchmark and the detailed captioning benchmark proposed by VideoRefer [96].

Method	#Params	Pos (%)	Neg (%)	Avg (%)
General VLMs:				
GPT-4o [54]	-	43.4	<u>79.6</u>	61.5
o1 [55]†	-	<u>46.3</u>	78.8	<u>62.5</u>
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**

Table 5: **Accuracies on detailed localized captioning in our proposed DLC-Bench.** DAM outperforms previous API-only models, open-source models, and region-specific VLMs on detailed localized captioning. <u>Underlined</u>: the second-best method. †: models with thinking mode.

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**

Table 6: **Detailed localized *video* captioning** on HC-STVG [71].

In Tab. 6, our proposed DAM achieves 19.8% relative improvement over the previous best on HC-STVG, including concurrent work VideoRefer [96]. In Tab. 7, the benchmark proposed by concurrent work VideoRefer [96], our proposed DAM surpasses the previous best in *both zero-shot and in-domain settings*, where zero-shot indicates not being trained on in-domain datasets derived from Panda-70M [17], which the benchmark also sources videos from.

Finally, we analyzed the performance of DAM in HD (hallucination detection) sub-task and found that DAM often predicts correct details not present in the reference caption. This indicates that the lower zero-shot performance on this sub-task is *not necessarily due to the hallucination of our model* but rather due to the missing details in the reference caption. We illustrate this further in App. C.2.

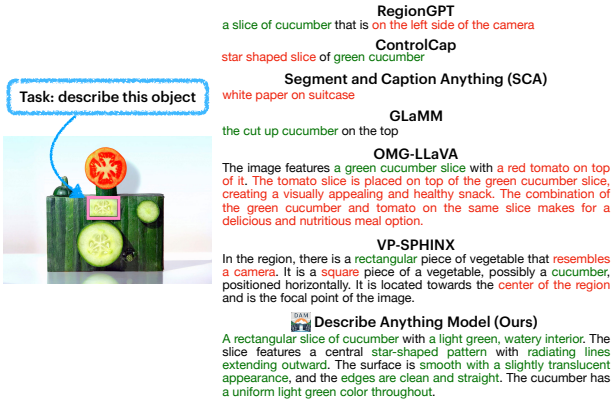


Figure 5: DAM generates detailed and localized descriptions, whereas prior works generate descriptions that are less precise. Green: correct description. Red: factual error or mislocalization.

6.2. Qualitative Results

Qualitative comparisons in Fig. 5 show that DAM excels in both accuracy and the level of details.

Detailed Localized *Video* Captioning. DAM describes user-specified objects in videos with localization from any frame. As shown in Fig. 6 **(a)**, DAM effectively describes objects under challenging conditions, such as motion and occlusion. *We offer more video examples in App. G.3.*

Controlling Description Granularity. As shown in Fig. 7, DAM allows control over the amount of details and length of descriptions with different prompts.

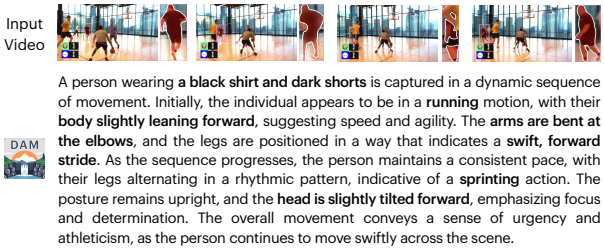
Zero-shot 3D Object Captioning. Our model can also describe objects in multi-view datasets such as Co3Dv2 [64], integrating information from multiple frames to provide coherent descriptions of 3D objects (Fig. 6(b)).

6.3. Ablations

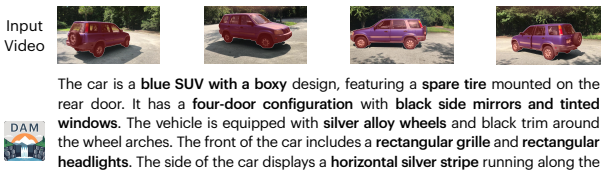
Visual Prompting. We analyze different prompting strategies and find that both localized inputs and contextual information are crucial for accurate descriptions. Using only the full image limits focus on specific regions (48.7%), while local crops improve detail but lose context (60.1%). Simply concatenating both performs poorly (42.4%). Adding cross-attention significantly improves performance (63.2%), and using focal crops further enhances results (65.4%). Our best approach, the **focal prompt**, integrates focal crops with cross-attention, achieving **67.3% accuracy** without increasing sequence length for the LLM.

Method	SC	AD	TD	HD†	Avg.
Zero-shot:					
Qwen2-VL-7B [81]	3.30	2.54	2.22	2.12	2.55
InternVL2-26B [20]	4.08	**3.35**	3.08	2.28	3.20
GPT-4o-mini [54]	3.89	3.18	2.62	2.50	3.05
GPT-4o [54]	4.15	3.31	**3.11**	2.43	3.25
Osprey-7B [95]	3.30	2.66	2.10	1.58	2.41
Ferret-7B [93]	3.20	2.38	1.97	1.38	2.23
Elysium-7B [80]	2.35	0.30	0.02	**3.59**	1.57
Artemis-7B [59]	3.42	1.34	1.39	2.90	2.26
DAM-8B (Ours)	**4.45**	3.30	3.03	2.58	**3.34**
In-domain:*					
VideoRefer-7B [96]	4.44	3.27	3.10	3.04	3.46
DAM-8B (Ours)	**4.69**	**3.61**	**3.34**	**3.09**	**3.68**

Table 7: **Performance on detailed localized video description on VideoRefer-Bench-D [96].** †: We provide analysis on hallucination scores (HD) in Sec. 6.1 and App. C.2. *: trained on in-domain VideoRefer-700k with regard to VideoRefer-Bench, both sourcing videos from Panda-70M [17].



(a) DAM reliably describes user-specified objects in videos, even under strong camera and object motion and heavy occlusion.



(b) DAM can describe objects in multi-view datasets, like this car from Co3Dv2 [64], by integrating information from multiple frames.

Figure 6: **DAM accurately describes user-specified regions in *videos* and *multi-view scenes* under challenging conditions.** More results presented in Fig. A.8.

Figure 7: **DAM offers multi-granular localized descriptions.**

Prompting	XAttn	#IT	Pos (%)	Neg (%)	Avg (%)
Full Image Only	No	196	32.1	65.4	48.7
Local Crop Only	No	196	43.5	76.6	60.1 (+11.4)
Full + Local Crop	No*	392	26.3	58.6	42.4 (-6.3)
Full + Local Crop	Yes	196	45.7	80.6	63.2 (+14.5)
Focal Crop Only	No	196	47.3	**83.6**	65.4 (+16.7)
Full + Focal Crop	Yes	196	**52.3**	82.2	**67.3** (+18.6)

Table 8: **Ablation studies across different visual prompts, cross-attention settings, and the number of image tokens.** Local crop denotes cropping without surrounding context. XAttn: cross-attention. #IT: number of image tokens. * indicates the full image and the crop are concatenated on the sequence dimension. **Bold**: Our proposed **focal prompt**.

Data	# regions	Pos (%)	Neg (%)	Avg (%)
LVIS	373k	34.0	72.6	53.3
+ additional datasets	602k	47.5	80.0	63.8 (+10.5)
+ SSL on 10% of SA-1B	1.38M	**52.3**	**82.2**	**67.3** (+14.0)

Table 9: **Our model benefits from diverse datasets generated by DLC-SDP.** Scaling the dataset in size and diversity significantly improves model performance, and SSL further enhances performance using widely available unannotated images.

Data Scaling. Expanding supervised datasets boosts performance, demonstrating the value of diverse regional data. Incorporating **semi-supervised learning (SSL)** with 10% of unannotated SA-1B images further improves accuracy to **67.3%**, showcasing our data pipeline's scalability.

Additional ablations, including training a model with the architecture from prior work on our data, comparing image-only vs image-video joint training, and ablating prompt augmentations, are in App. E.

7. Discussions and Conclusion

We introduced DAM, a model for detailed localized captioning in images and videos, balancing local detail and global context through a focal prompt and localized vision backbone. We developed DLC-SDP, an SSL data pipeline leveraging segmentation datasets and unlabeled web images for high-quality captions. We also proposed DLC-Bench, a benchmark using attribute-based evaluation to overcome limitations of reference-based scoring. DAM achieves SOTA in 7 benchmarks in multi-granular regional captioning. We present additional discussions in App. C.

A. Challenges in Generating Detailed Localized Descriptions with Off-the-Shelf VLMs

Although cutting-edge Vision-Language Models (VLMs), such as GPT-4o [54] and LLaVA [46–48], excel at generating global-level image descriptions, producing detailed *localized* image captions remains an open problem. Specifically, these VLMs only take in RGB images along with text prompts and do not allow users to accurately specify regions of interest.

While users could employ text to localize the object to be described, this approach is often cumbersome and inefficient, requiring precise referring phrases that may still be difficult for the VLM to interpret. This can lead to mislocalization of the intended object, as illustrated in Fig. A.1(a).

The required effort for both the user and the model can be significantly reduced if the user is allowed to specify the region directly using a representation in 2D coordinates that the model can understand. With this idea in mind, we focus on *generating detailed localized descriptions* by enabling users to specify a region in an image for the model to describe in detail. Since spatial representations such as points and boxes can be converted into masks using SAM [32] and SAM 2 [62], we concentrate on regions specified by mask inputs.

A first attempt to address this problem with existing VLMs is to reduce the task to global image captioning by presenting only the region to the VLM through masking or cropping, as shown in Fig. A.1(b). While this forces the VLM to focus solely on the specified region, freeing users from the burden of expressing localizations as phrases, the lack of contextual information makes the task much more challenging and often confuses the VLM. This confusion can prevent the model from correctly identifying the object, let alone providing detailed descriptions of its parts. In more extreme cases, the model may even refuse to caption the region due to insufficient information in the cropped or masked image. Therefore, generating detailed localized captions requires more than just the local region.

An alternative approach to prompt existing off-the-shelf VLMs for localized descriptions is to overlay markings such as points, scribbles, contours, and alpha masks on the image [91, 92], as shown in Fig. A.1(c). However, these markings may blend into the object or the background in highly complex scenes, making them unrecognizable to the VLMs. This issue is especially common for small objects that are not the main focus of the scene. Furthermore, the markings may render the image out-of-distribution, confusing the VLMs and disrupting the quality of output that they were originally capable of generating.

The exploration above highlights a conflict between the precision of localization and the availability of context. On one hand, we want the model to accurately focus on a specific region without mentioning other regions, such as other objects or the background. On the other hand, the model needs to leverage contextual information to correctly identify the object in the region of interest. This conflict makes it very difficult for current VLMs to produce high-quality localized descriptions.

Our proposed model overcomes this challenge by taking the localization as a *separate* input in 2D space. This approach has the advantage of making the localization more explicit for the VLMs to parse while keeping the image within its original distribution, thus preventing the model from being distracted by the markings. This technique leads to accurate localization even in complex scenes, as illustrated in Fig. A.1(d). Note that since Fig. A.1(d) mainly focuses on explaining the design choices of inputting mask inputs to the model, focal prompting is included as a part of the model and is omitted in this figure for simplicity. We refer readers to Fig. 3 for illustrations on focal prompting.

B. Evaluation Benchmarks

Our DAM is designed to perform well at *localized image and video captioning* across *multiple granularities*, including keyword, phrase, and detailed captions. Therefore, we evaluate and achieve SOTA in 7 in-domain and zero-shot benchmarks:

1. The LVIS open-class keyword-level benchmark in Tab. 2.
2. PACO open-class keyword-level benchmark (including object and parts as regions) in Tab. 2.
3. Flickr30k Entities phrase-level benchmark in Tab. 3.
4. Ref-L4 detailed captioning benchmark in Tab. 4.

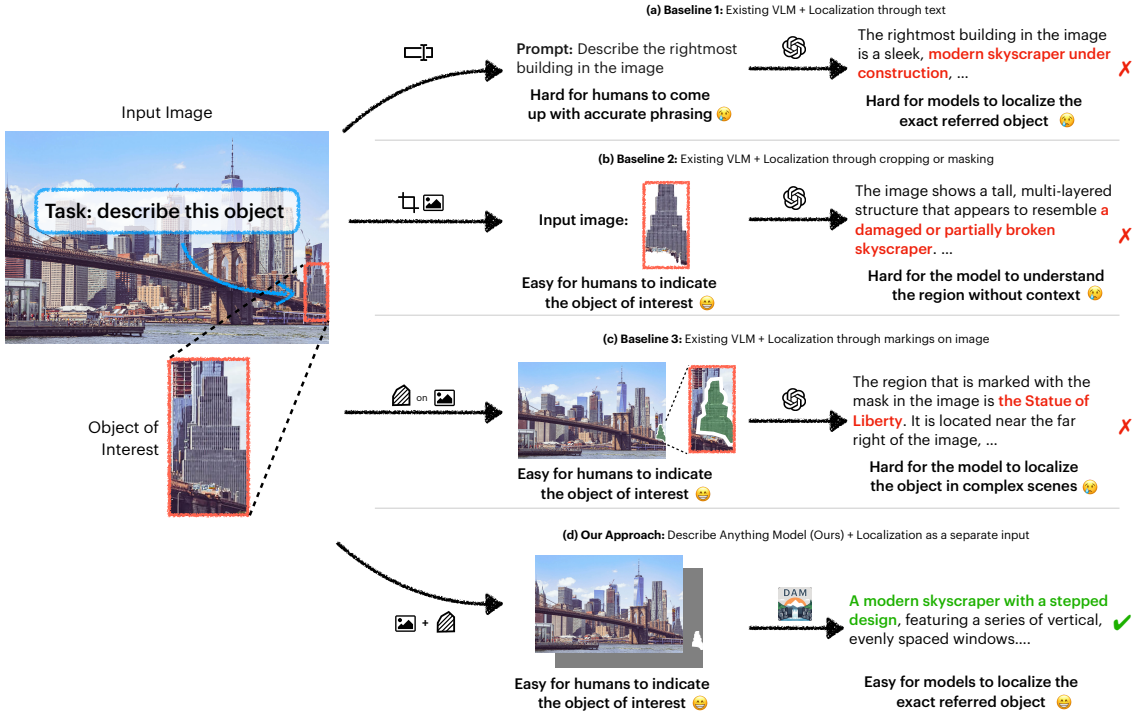


Figure A.1: **Existing Vision-Language Models (VLMs) do not perform well in generating localized descriptions.** (a) to (c) demonstrate several ways to prompt existing VLMs, but none achieves satisfactory performance, leading to the need for a new method that is capable of providing detailed and localized descriptions. In (d), we propose a model that accepts the condition in a separate form of input, making it easy for users to specify the object of interest and for the models to accurately localize the referred object. Note that our focal prompt, proposed in Sec. 3, is considered part of the Describe Anything Model and is not shown in the figure for simplicity.

5. Our proposed DLC-Bench detailed localized captioning benchmark in Tab. 5.
6. HC-STVG detailed video captioning benchmark in Tab. 6.
7. VideoRefer detailed video captioning benchmark in Tab. 7.

We offer an explanation for each setup.

B.1. Keyword-level Localized Captioning Benchmarks

Open-class keyword-level localized captioning benchmarks, proposed in [95], require the model to output keywords containing the object and part entities to describe the region. In contrast to closed-class keyword-level localized captioning, which constraints the model output to several choices provided, open-class keyword-level localized captioning takes free-form text outputs from the model. The evaluation results are in Tab. 2.

1. For LVIS [29], this involves predicting the class name as a keyword, given the segmentation mask of an object. A typical class name ranges from one word to four words.
2. For PACO [60], this involves predicting the class name of an object in the mask if the mask contains a full object, or the object name and the part name if the mask contains an object part. This is especially challenging because it would require the model to understand nuances between full objects and object parts.

B.2. Phrase-level Localized Captioning Benchmarks

Phrase-level localized captioning task requires the model to output a phrase containing a brief description for each the region that includes object identification and attributes typically within a few words. The metrics typically used in phrase-level benchmarks are CIDER, METEOR, BLEU, ROUGE_L, and SPICE

[2, 5, 43, 56, 79]. We refer these metrics as short captioning metrics, as opposed to metrics from LLM-based evaluations that support evaluating detailed captions.

We perform zero-shot evaluation on the grounded phrases in Flickr30k Entities [58], where our model is not trained on the entities annotated in the training split of Flickr30k Entities. Results are in Tab. 3.

B.3. Detailed Localized Captioning Benchmarks

Detailed localized captioning task requires the model to output a detailed description for each the region with the length spanning from a long sentence to multiple sentences.

1. We perform zero-shot evaluation on detailed captions in the Objects365 [67] split of Ref-L4 [14] since we do not train on Objects365 dataset. We evaluate the prediction quality by computing short captioning metrics and CLAIR [13] score against the reference captions in the dataset. We use CLAIR to evaluate raw detailed outputs, while we summarize both the prediction and ground truth with GPT-4o-mini [54] before evaluation with short captioning metrics. No ground truth or reference captions are provided to GPT-4o-mini, with the LLM setting exactly the same for all models for fairness. Results are in Tab. 4.
2. We evaluate our model with DLC-Bench, our proposed benchmark for fine-grained region-based captioning. This evaluation is also zero-shot. We present details about our benchmark in App. D. Results are in Tab. 5.

B.4. Detailed Localized Video Captioning Benchmarks

1. We conduct evaluation on HC-STVG [71], a spatial-temporal video grounding dataset with detailed captions used in prior and concurrent work [59, 96]. Following prior work [59], we evaluate the quality of localized captions with CIDER, METEOR, BLEU, ROUGE_L, and SPICE [2, 5, 43, 56, 79]. Results are in Tab. 6.
2. We also perform evaluation on the detailed localized video description benchmark in VideoRefer-Bench proposed by concurrent work [96]. GPT-4o is used to provide four dimensions of scores on a scale of 1 to 5. The four dimensions are Subject Correspondence (SC), Appearance Description (AD), Temporal Description (TD), and Hallucination Detection (HD). Zero-shot setting indicates that our model is not trained on Panda-70M [17], the dataset that VideoRefer-Bench sources the videos from. In-domain setting indicates mixing the detailed caption subset of VideoRefer-700k, which is also curated from Panda-70M [17], into our training data. Results are in Tab. 7.

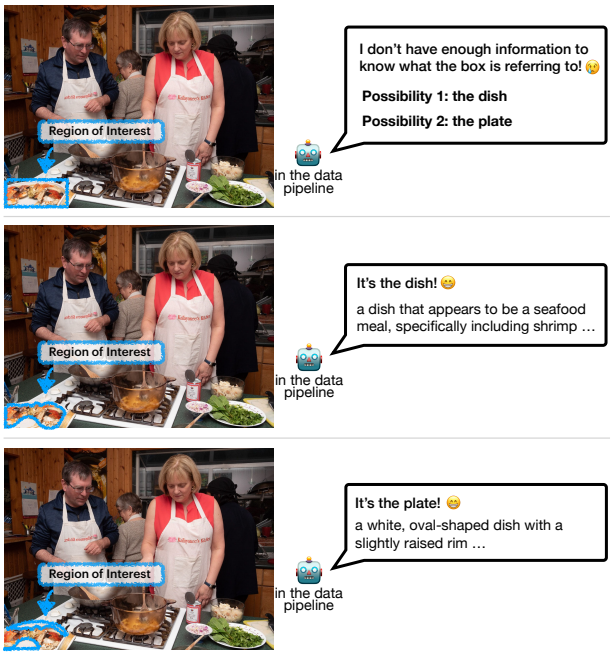


Figure A.2: **Caveats for using boxes to indicate region of interests.** Top: Using a box to indicate the region of interest leads to ambiguity. Middle and Bottom: Switching to a mask representation leads to more specific referring and correct descriptions.

C. Discussions

C.1. The Caveats of Using Referring Boxes in Data Pipeline

Caveats exist when boxes are used to refer to regions in the data pipeline. As shown in Fig. A.2, boxes can be ambiguous in terms of what they are referring to, causing uncertainty for the VLM that we use in our data pipeline. In contrast, masks are much more specific in terms of the region that it is referring to. This motivates us to use manually annotated masks in existing segmentation datasets rather than bounding boxes in order to curate high-quality data for DLC with little referring ambiguity. We additionally take

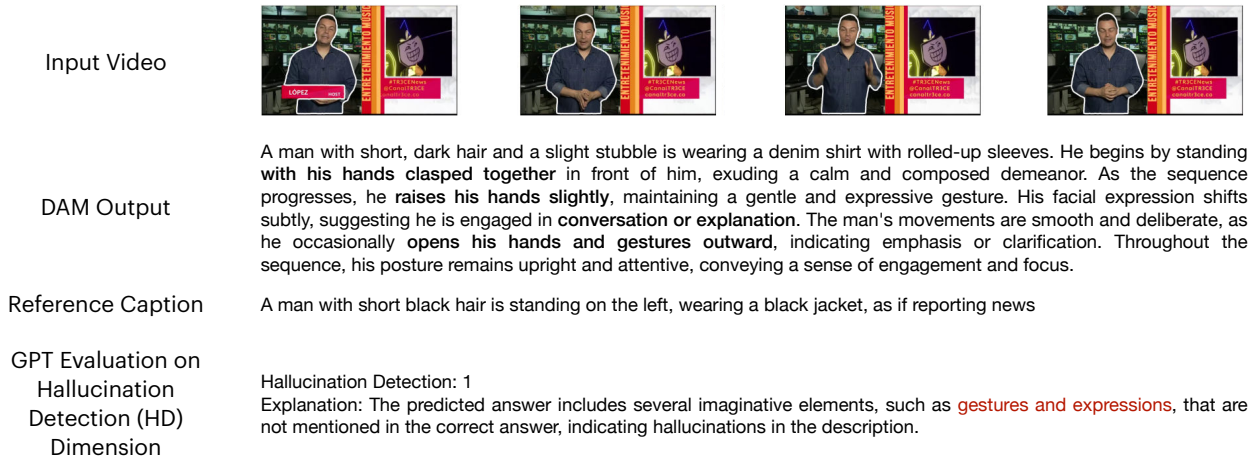


Figure A.3: **The pitfall of using reference captions for caption evaluation.** Evaluation benchmarks based on reference captions may incorrectly treat correct details in the predicted caption as hallucination. Since the GPT evaluator relies solely on the ground truth caption without viewing the video, it mistakenly flags gestures and expressions as hallucinations, resulting in a low score. However, the evaluation is invalid since the predicted details are correct.

in manually annotated keywords (*e.g.*, class names, part names, entities) in the datasets for regions we are annotating in our data pipeline to further reduce the ambiguity and potential confusion for our VLM in the data pipeline.

C.2. The Pitfall of Using Reference Captions in Benchmarks

As discussed in Sec. 5 and 6.1, caveats exist for using a "ground truth" reference caption for benchmarking localized descriptions. Specifically, since such a reference caption is hardly comprehensive and may not contain all the details about the region of interest, the metrics from the benchmark will treat the correct details in the caption prediction about the region of interest that are not mentioned in the ground truth reference caption as hallucinations. This discourages the model from generating detailed captions.

We analyzed the performance of our method in HD (hallucination detection) sub-task in VideoRefer-Bench [96] and found that our model often predicts correct details that are not present in the reference caption. Specifically, the example in Fig. A.3 shows this phenomenon. While our model's prediction includes appearance and motion details about the change of the person's gesture and expression, such details are not mentioned in the reference caption in the dataset. Since the GPT evaluator does not see the video and uses the ground truth caption as the only source of information, it incorrectly believes that the gestures and expressions are hallucinations and gives our caption a very low score for the hallucination detection dimension. However, the evaluation is not valid, as our model is correct in the descriptions about the gestures and expressions.

This indicates that the lower score on this sub-task is *not due to the hallucination of our model*, but rather due to the missing details in the reference caption and the fact that our model, evaluated in a zero-shot setting, does not have awareness for what types of details are preferred by or included in the reference caption.

C.3. Failure Cases

We show two failure cases of DAM in Fig. A.4. In Fig. A.4(a), DAM misrecognizes the frog-shaped slipper to be a frog. In Fig. A.4(b), DAM describes the person as pulling the body upward. We expect these errors to be mitigated by broader data coverage.

C.4. Potential Limitations

DAM is only trained for multi-granular localized captioning, especially for detailed localized captioning (DLC) and is not specifically optimized for other general vision-language tasks. However, DAM is designed for in-depth analysis for the task of multi-granular image and video localized descriptions rather than for breadth for general vision-language understanding,

DAM (Ours)

A **bright green frog** with a smooth, glossy texture. The frog has a rounded head with prominent, slightly raised eyes. Its body is streamlined, tapering towards the tail, and the limbs are short and sturdy.

(a) DAM might still misrecognize a region and output an incorrect description. For example, it misrecognizes the frog-shaped slipper to be a frog.

Input Video

DAM (Ours)

A person wearing a sleeveless white top and blue shorts is engaged in a dynamic exercise routine. The individual is equipped with a pair of red and white athletic shoes, which provide stability and support during the activity. The person is holding onto a bar or handle with both hands, suggesting a pull-up or similar upper-body exercise. The posture is upright, with the back slightly arched, indicating the exertion of muscles during the movement. The legs are bent at the knees, with feet firmly planted on the ground, providing a stable base. Throughout the sequence, the person maintains a consistent rhythm, **pulling the body upward** and then lowering it back down in a controlled manner. This repetitive motion showcases the individual's strength and endurance, as they continue the exercise with focus and determination.

(b) DAM might also be confused by the combination of the object motion and the camera motion. In this example, it makes the mistake of describing the person as pulling the body upward.

Figure A.4: Failure cases for our proposed DAM.

which justifies the design choice.

C.5. Computational Efficiency

DAM incorporates our proposed localized vision encoder, which differs from the SigLIP [97] vision encoder used in [44] by adding two key components: *patch embedding layers* for encoding the mask and *cross-attention blocks*. Importantly, these components do not alter the dimensions or sequence length of the vision features passed to the large language model, ensuring that the parameter count and computational efficiency of the large language model are unaffected. Since the vision encoder represents only a small fraction of the total parameters and computational operations, the overall increase in FLOPs and parameter count remains marginal, maintaining the model's efficiency.

To be more specific, unlike prior works that derive regional features from *image features* for each region, the regional feature used in our approach comes directly from *a global and a focal view of the input image*, with cross-attention enhancing the focal representation. This design is justified as the vision encoder is much smaller than the LLM (400M vs. 3B/8B parameters), with minimal latency impact (*0.06s compared to 1.49s for 3B LLM* as measured in our pipeline). This overhead is outweighed by the benefits of preserving fine details that global image features miss as indicated in Tab. 8), especially for small regions. Finally, DAM 3B outperforms much larger models in challenging (Tab. 5), showing our efficiency.

C.6. Training Data

In addition to the details in data annotation presented in App. H.1, we discuss the training data of our work in this section and present a comparison with recent works. Compared with recent work Ferret [93] which used 1.1M *unreleased* samples and RegionGPT [28] which used 1.5M *unreleased* samples,

we train our model on a comparable amount of data (1.5M samples). However, we obtain much better performance (Tab. 5), which shows the effectiveness of DAM.

C.7. Performances of Baseline Models on DLC-Bench

Interestingly, region-specific VLMs often perform on par or worse than generic VLMs. This is likely because many are trained on datasets with short regional captions, leading them to produce brief, phrase-level descriptions. Even when prompted for longer descriptions [28, 100], these models tend to include irrelevant details about the background, speculations, or hallucinations, due to insufficient regional information. Providing crops instead of full images leads to mixed results for different region-specific VLMs since these models are not designed to describe regions in crops.

D. Details for DLC-Bench

Image and Instance Selection. We leveraged a subset of the Objects365 v2 [67] validation set, which was manually annotated with segmentation masks in [27], for image and instance selection. We collected a set of 892 challenging questions from this subset, each containing one object of interest. Each question is manually inspected, and questions with ambiguous or unclear answers are filtered out. To maintain the integrity of our benchmark, we conducted de-duplication to ensure that no images used in the benchmark were present in our training dataset for detailed localized captioning.

Positive Question Generation. For each masked region, we prompted an off-the-shelf Visual Language Model (VLM) to generate a list of parts. Subsequently, for the whole object and each part, we asked the VLM to generate a list of properties covering aspects such as color, shape, texture, materials, and size. Each property is stored in the form `([object name], [part name], [property name], [property value])`. For example, if the masked region is a corgi, the VLM could describe the brown fur of the corgi as `(corgi, fur, color, brown)`.

We used this list of properties as a starting point for manual curation. We then manually added significant properties that the VLM missed, revised inaccurate properties, and removed hallucinated or ambiguous properties from the VLM outputs. Finally, we turned these properties into questions that test whether a description accurately covers the property.

Negative Question Generation. We targeted mislocalization and hallucination, which are two types of negatives (*i.e.*, cases in which a property or an object should not be included in the descriptions). Specifically, for mislocalization errors, we prompted the VLMs to generate a list of objects in the image that are not in the masked region. We also prompted the VLMs to generate a list of parts that are commonly associated with the object type of the masked region but are not present or visible in the masked object in the image (*e.g.*, the head of a corgi if it is occluded and thus not included in the masked region).

To avoid biasing towards one specific off-the-shelf VLM, we leveraged multiple VLMs for different instances to generate initial positives and negatives. Specifically, we annotated 34 regions using GPT-4o [54], 35 using Gemini 1.5 Pro [75, 76], and 31 using Anthropic Claude 3.5 Sonnet [72] for the initial property generation. We used the same image prompting method for all VLMs as we did when prompting the VLMs in the first stage of our data pipeline.

Note that the choices for each question are mutually exclusive, which ensures one option is always valid and leaves no room for two options to be true at the same time.

Scoring Mechanism. Our evaluation methodology involves scoring the models based on their ability to include correct details and exclude incorrect or irrelevant information.

To evaluate a model like DAM for its ability to output detailed localized captions, we first prompt the model to generate descriptions for each of the masked instances. Then, instead of directly asking our model to provide answers to these questions, we prompt a text-only LLM, Llama 3.1 8B [25], to serve as a judge to rate the localized descriptions according to the positive and negative questions.

For each model-generated description, we apply the following scoring rules:

- **Positive Scoring:** For each positive question, if

(a) An example region in the benchmark for the model to describe

A white electric stove with a smooth, flat cooktop featuring two visible black burners. The control panel is located at the back, with a black surface and a few visible knobs. The stove has a slightly raised backguard.

(b) An example description to be evaluated

Q: Which of the following is applicable to the description?
Choices:
A. The control panel or the stove is not mentioned. (score: 0)
B. The location of the control panel is mentioned in the description but is not back. (score: -1)
C. The location of the control panel is mentioned in the description and is back. (score: 1)
D. The location of the control panel is not mentioned, but the control panel of the stove is mentioned. (score: 0.5)

(c) An example positive question in DLC-Bench

Q: Which of the following is applicable to the description?
Choices:
A. The induction cooktop surface of the stove is mentioned in the description. (score: -1)
B. The induction cooktop surface of the stove is not mentioned in the description. (score: 1)
C. The stove is not mentioned in the description. (score: 0)

(d) An example negative question in DLC-Bench

Figure A.5: **An example from DLC-Bench for detailed localized captioning. (a)** The process begins by prompting a model to describe a specified region within the image. The resulting description is then evaluated using a text-only LLM as a judge that rates each response by answering positive and negative questions. **(b)** shows an example description to be evaluated. **(c)** Positive questions are designed to test whether the model correctly identifies specific details within the described region. The model receives points for accurate details and is penalized for factual errors. The bold option (option C) indicates that the LLM judge believes that option C is applicable, allowing the model to get a point for this example positive question. **(d)** Negative questions ensure the model refrains from mentioning irrelevant or nonexistent details. Mislocalization or hallucination results in penalties to prevent false positives. The bold option (option B) indicates that the LLM judge believes that option B is applicable, allowing the model to get a point for this negative question.

the description correctly includes the specified detail, the model receives a point. To prevent models from artificially inflating their scores by generating excessively long descriptions and guessing details, we penalize incorrect details and discourage models from including uncertain or erroneous content. If the detail is mentioned but incorrectly (*e.g.*, wrong color), a penalty of one point is applied. No point is awarded if the description does not mention the detail. Partial points (0.5 points) are awarded for answers that are partially correct but insufficiently detailed. Note that the model gets positive points only when the object recognition is correct, as the correctness of the details depends on the correctness of the overall region recognition. We present a positive example in Fig. A.5(c).

- **Negative Scoring:** For each negative question, if the description appropriately excludes the incorrect or irrelevant detail, the model gets a point. If the description includes the detail, indicating mislocalization or hallucination, a penalty is applied. The model gets zero or negative points when the object recognition is incorrect, since otherwise a caption that is random and completely off could get high scores on the negative questions. We present a negative example in Fig. A.5(d).

The positive (negative) score for a model is the sum of points for positive (negative) questions, normalized by the maximum possible score to yield a percentage for comparison. We also average the positive and negative scores to obtain an overall score, which represents the model's overall capability in detailed localized captioning.

We present an example from DLC-Bench in Fig. A.5. The example region in Fig. A.5(a) features a stove with coil burners. An example description of the region is presented in Fig. A.5(b). For the example positive question in Fig. A.5(c), the LLM judge selects option C, as the caption correctly mentions that the control panel is at the back, allowing the model to get a point for this positive question. For the negative question in Fig. A.5(d), the LLM judge selects option B, as the caption correctly indicates that it is not an induction

cooktop, allowing the model to get a point for this negative question.

Evaluation Setting. For our models, we follow our inference setting described in App. H.

E. Additional Ablation Studies

Model Architecture with the Same Training Data. A model's performance is largely due to two factors: model architecture design and training data. Since both factors differ for different models, it is hard to compare the effectiveness of different model architectures head-to-head.

To this end, we compare our model architecture against VP-SPHINX [45], the strongest prior baseline in most benchmarks that we tested on. By continuously training a VP-SPHINX model [45] on our data after pre-training on the originally proposed datasets. This is a fair comparison since our method is also fine-tuned from a pretrained VLM, VILA-1.5 [44], with two stages of training prior to training on our region-specific dataset.

As shown in Tab. A.1, our model architecture achieves much better performance on detailed localized captioning benchmark DLC-Bench with trained on the same dataset from our proposed data pipeline. This justifies that our proposed focal prompt and localized visual backbone are able to provide more detailed features compared to just the global image features extracted by a vision encoder on the full image with a regional referring feature, as employed in [45].

Prompt Augmentation. We compared variants of our model with and without prompt augmentation. As shown in Tab. A.2, incorporating prompt augmentation slightly degrades our model's performance on the positive questions in our benchmark. We hypothesize that despite introducing variations in the prompts and enhancing the model's instruction-following capabilities, prompt augmentation creates a mismatch between the prompts used during training and those used during evaluation (as we always use the same prompt for evaluation, which is detailed in App. H.3). Since the prompt used during evaluation might not be the same as the prompt used in training, the model may also occasionally reference other tasks from our mixing dataset ShareGPT-4V for the length of outputs. This may cause the model to produce outputs that are not as detailed as when it is trained exclusively with the original prompt. Importantly, the model's performance on the negative questions remains unchanged, indicating that prompt augmentation does not lead to hallucinations or mislocalization.

Despite the slight degradation of the performance in the benchmark (0.6% in the overall accuracy), we observed that prompt augmentation improves instruction-following capabilities when prompts include additional instructions, particularly those specifying requirements on the length of the outputs. Therefore, we default to using the model without prompt augmentation in our benchmark evaluations, including ablations. In contrast, we employ the model with prompt augmentation in the qualitative evaluations.

Image-only Training vs Image+Video Joint Training. We also compared our image-only DAM with DAM with image + video joint training in Tab. A.3. We show that our model with image-video joint training slightly outperforms our model with image-only training on detailed localized image captioning. Note that for this ablation study, we keep the model size the same and use the 3B model for both image-only training and image-video joint training. We use image-only training as the default option for results in our benchmark for simplicity.

F. Additional Quantitative Results

Set-of-Marks Prompting. We present a comparison with baseline VLMs that use Set-of-Marks (SoM) prompting [91] in Tab. A.4. SoM leads to degraded results compared to the prompt engineering method used in stage one of our data annotation pipeline. This is mostly because the marks proposed by SoM blend in with the object or the background in complex scenes. They might also mask out some part of the object, which interferes with the model's understanding capabilities. Therefore, for fair comparisons, we use the same prompt engineering method as we use in stage one of our data annotation pipeline in our main result in Tab. 5. Importantly, region-specific VLMs, including DAM, have predefined ways of encoding regional inputs, making SoM inapplicable to these models.

	VP-SPHINX Arch	Our Arch
Avg (%)	50.2	**63.8**

Table A.1: **Ablations on architecture design compared to our strongest baseline VP-SPHINX [45].** We trained a model with VP-SPHINX [45] architecture on our curated DLC data from various segmentation datasets. The results on DLC-Bench indicate the advantages of our model architecture that allows detailed localized features to be presented to the LLM for DLC.

Prompt Augmentation	Pos (%)	Neg (%)	Avg (%)
No	**52.3**	**82.2**	**67.3**
Yes	51.3	**82.2**	66.7

Table A.2: **Comparison of performance of DAM with and without prompt augmentation.** Prompt augmentation has minimal effect on DAM's performance on DLC-Bench. While descriptions generated by the model may occasionally be less detailed, leading to a slight decrease in the performance on positive questions, we observed that prompt augmentation enhances instruction following when prompts include specific guidelines, such as length constraints. We use the model without prompt augmentation with our benchmark, including ablations, by default.

Setting	Pos (%)	Neg (%)	Avg (%)
Image-only Training	52.3	82.2	67.3
Image+Video Joint Training	**52.4**	**85.4**	**68.9**

Table A.3: **Comparison of performance of our image-only DAM and DAM trained with both localized image description task and localized video description task.** Joint training benefits generating high-quality localized image descriptions compared to image-only training.

G. Additional Qualitative Results

G.1. Detailed Localized Image Captioning

In Fig. A.7, we present additional examples from LVIS [29] to show our model's strong performance on detailed localized image captioning.

Our model demonstrates robust localization and region understanding capabilities. In the first example, it accurately describes the sofa cushion without mentioning the dog that is outside the masked region. In the second example, it correctly identifies the roller blind, which would be challenging to recognize based solely on a local crop without context. In the third example, the model provides a detailed description of the giraffe without referencing the birds perched on it, as they fall outside the masked region. These examples highlight our model's precise localization abilities and its effectiveness in perceiving regional details with contextual understanding.

G.2. Zero-shot QA Capabilities

Although not trained on any regional QA datasets, DAM surprisingly exhibits emerging zero-shot capabilities on regional QA.

In Fig. A.6, we show examples of our model performing zero-shot QA. DAM is able to identify properties of objects in the masked regions. For example, it is able to identify the color of the clothing, the material of the stick, and the textural pattern of the fish in the first three examples. DAM is also capable of performing object recognition for a region in the image, identifying the strawberry in the last image.

G.3. Detailed Localized Video Captioning

We present more examples for detailed localized video captioning in Fig. A.8 and Fig. A.9. Our model can describe objects in videos with large object motion and camera motion. DAM can also identify stationary objects by indicating that they are stationary in the description.

Method	#Params	Pos (%)	Neg (%)	Avg (%)
API-only General VLMs:				
GPT-4o (SOM) [54]	-	5.0	29.2	17.1
o1 (SOM) [55]†	-	0.8	28.0	14.4
Claude 3.7 Sonnet (SOM) [73]†	-	0.5	40.2	20.4
Gemini 2.5 Pro (SOM) [74, 75]†	-	13.2	65.0	39.1
Open-source General VLMs:				
Llama-3.2 Vision (SOM) [25]	11B	16.8	40.4	28.6
Llama-3 VILA1.5 (SOM) [44]	8B	0.6	0.6	0.6
InternVL2.5 (SOM) [20, 21, 84]	8B	8.6	28.6	18.6
LLaVA v1.6 (SOM) [46–48]	7B	2.2	3.8	3.0
Qwen2.5-VL (SOM) [77, 81]	7B	8.5	27.2	17.8
VILA1.5 (SOM) [44]	3B	-0.4	15.4	7.5
DAM (Ours)	3B	**52.3**	**82.2**	**67.3**

Table A.4: **Additional results with existing general VLMs using Set-of-Mark (SoM) prompting [91].** The results are accuracies on detailed localized captioning in DLC-Bench. Compared with results in Tab. 5 which are obtained with the same prompt engineering as we use in the stage 1 of our data pipeline, SoM leads to degraded quality. In this comparison, the advantages of our method, compared with prior baselines, are much larger. Negative numbers are due to penalties from factual errors. Note that region-specific VLMs, including our proposed DAM, have predefined ways of inputting regions, and thus SoM prompting is not applicable to these models. †: models with thinking mode.

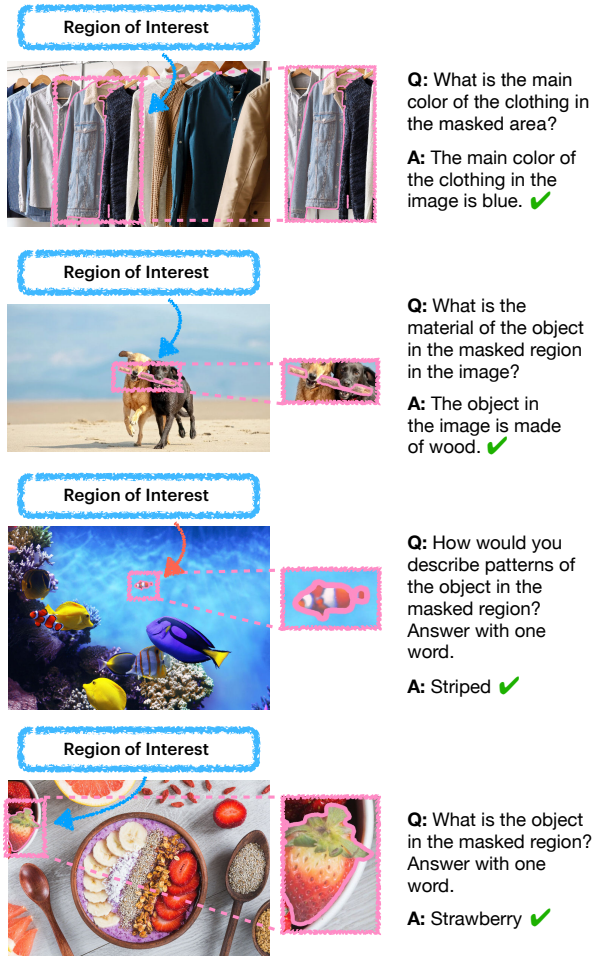


Figure A.6: **Emerging zero-shot QA capabilities.** DAM could answer questions about regions in an image, showcasing capabilities such as object recognition and property identification.

G.4. Qualitative Comparisons with Strong Baselines

Detailed Localized Image Captioning. We also present qualitative comparisons with GPT-4o [54] and our strongest open-weight baseline VP-SPHINX [45] in detailed localized image captioning in Fig. A.10.

In both examples, GPT-4o could not correctly recognize the objects in the masked regions, providing only vague descriptions. VP-SPHINX, while better than GPT-4o, still struggles with accurate object recognition and detailed descriptions. In the left image, VP-SPHINX incorrectly describes a group of seals when the masked region contains only one seal. In the right image, VP-SPHINX identifies the towel but provides minimal detail, missing key attributes like its color and texture.

Our model outputs detailed and high-quality descriptions of the seal and the towel. This improvement stems from our model's design which enables the fusion of object-specific information with broader contextual understanding.

Detailed Localized Video Captioning. We present comparisons with three strong video understanding models, GPT-4o [54], Qwen2.5-VL [77], and recent work VideoRefer- [96], in detailed localized video captioning in Fig. A.11. In the top example, we observed that GPT-4o struggles to interpret the cow's movements. Similarly, Qwen2.5-VL-7B incorrectly perceives the cow as stationary. VideoRefer-7B provides minimal motion and appearance details. In contrast, our 8B model accurately identifies the motion of the cow, providing more detailed information about it.

In the bottom example, GPT-4o misidentifies the object, mistakenly assuming the animal is transforming into a wolf or a pig. Meanwhile, Qwen2.5-VL-7B believes only the sheep's head is moving. VideoRefer-7B recognizes that the sheep is moving but provides little detail about the appearance of the sheep. In contrast, our model correctly identifies the animal in the

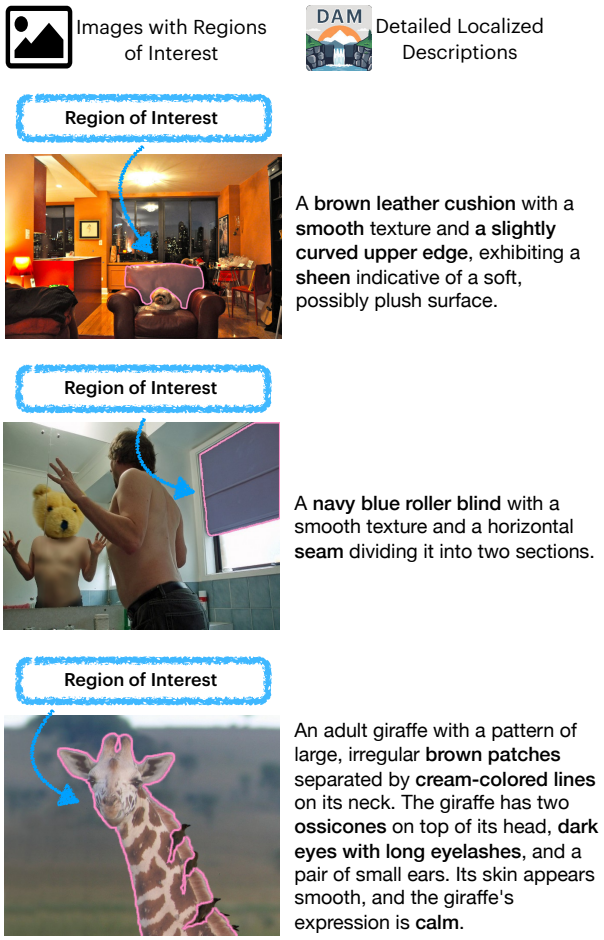


Figure A.7: **Additional results from LVIS [29] demonstrating DAM's detailed localized image captioning capabilities**. Our model exhibits robust region understanding and localization across diverse scenarios. It produces precise descriptions of objects within masked regions while successfully identifying challenging details like the roller blind in the second example through effective use of contextual cues.

masked region as a sheep throughout the video and accurately recognizes its full movement, providing details about its motion and appearance.

H. Implementation Details

H.1. Data Annotation Pipeline

Stage 1. We annotate four existing instance and semantic segmentation datasets for detailed localized descriptions. We use off-the-shelf VLMs for region annotations, with 603k regions across 202k images with detailed localized descriptions in total in stage 1. For the model variant used in PACO [60] open-class dataset evaluation, we additionally merged in 81k annotated instances from PACO [60] to improve its part description capabilities, leading to 684k annotated regions, as detailed in Tab. A.5. To prompt a VLM to output detailed localized descriptions, we input a cropped image and a masked image. While the cropped image allows coarse localization and provides high token density per pixel for clear descriptions, the masked image helps localize the object of interest when there are multiple instances with the same category. The category name is also provided in the text prompt, relieving the model from having to identify the object without the context from the image. We present the prompt for data annotation in Tab. A.7.

Stage 2. We annotate 10% of SA-1B through self-

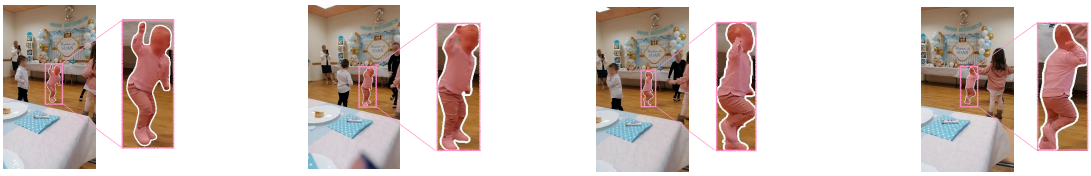
Dataset	# Images	# Regions
Stage 1:		
LVIS [29]	90,613	373,551
Mapillary Vistas v2.0 [53]	17,762	100,538
COCO Stuff [11]	28,365	32,474
OpenImages v7 [33, 35]	64,874	96,006
PACO [60]	24,599	81,325
Stage 2:		
SA-1B (10%)	592,822	774,309
Total	**819,035**	**1,458,203**

Table A.5: **Dataset statistics across stages with total images and regions for training detailed localized image captioning.** In stage 1, we annotated 684k regions across 226k images from existing instance and semantic segmentation datasets. In stage 2, we perform SSL on 10% of SA-1B images without using the masks provided by the datasets, resulting in 774k regions across 593k images. In total, we annotated 1.46M regions across 819k images with detailed localized descriptions. This diverse and high-quality dataset is the key to our model's performance. Note that due to filtering the number of instances and images are lower than the number of instances and images in the original dataset.

Dataset	# Videos	# Regions
SA-V [62]	36,922	93,969

Table A.6: **Dataset statistics across stages with total videos and regions for training detailed localized *video* captioning.** We label 94k regions across 37k videos from SA-V dataset [62] for detailed localized video captioning. Note that each region indicates an instance across multiple frames in the video.

Input Video

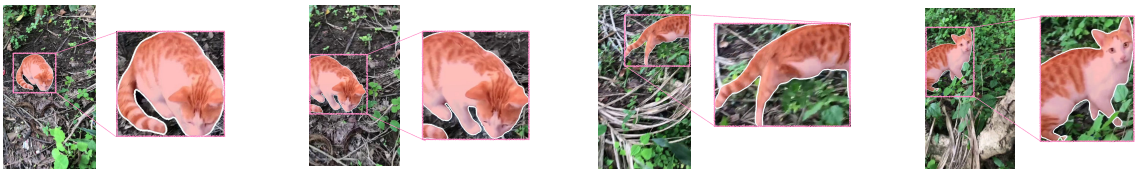


DAM

A young child, dressed in a light-colored long-sleeve shirt and beige pants, is captured in a sequence of dynamic movements. The child appears to be in mid-motion, possibly dancing or playing, with one arm raised and the other slightly bent. The child's posture suggests a lively and energetic demeanor, with a slight forward lean indicating movement. The legs are positioned in a way that suggests a step or a jump, with one foot slightly lifted off the ground. The child's head is tilted slightly upward, adding to the sense of motion and engagement. Throughout the sequence, the child maintains a consistent posture, with the arms and legs moving in coordination, conveying a sense of playful activity.

(a) DAM accurately describes fine-grained details for the masked region.

Input Video

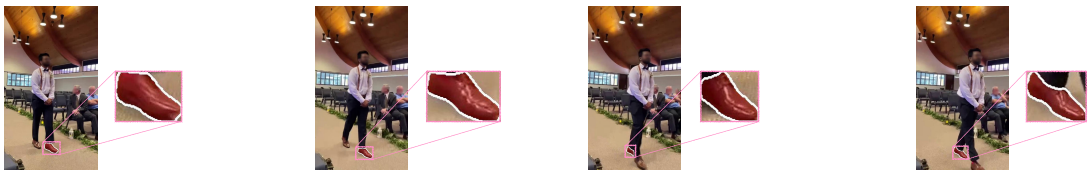


DAM

A cat with a predominantly white coat and distinct orange patches is seen in a series of fluid movements. Initially, the cat is crouched low, its body compact and poised, suggesting a state of alertness or curiosity. As the sequence progresses, the cat begins to rise, its body elongating and stretching, indicating a transition from a crouched position to a more upright stance. The cat's ears are perked, and its eyes are focused, suggesting attentiveness to its surroundings. The tail, initially curled, extends outwards, providing balance and a sense of readiness. The cat's movements are smooth and deliberate, showcasing its agility and grace. As it continues to rise, the cat's posture becomes more upright, and its head turns slightly, as if scanning the environment or preparing to move in a new direction. The sequence captures the essence of a cat's natural curiosity and fluid motion, highlighting its ability to navigate its environment with ease and elegance.

(b) DAM is still able to output high-quality descriptions in scenes with fast camera and object motion.

Input Video



DAM

A polished brown leather shoe, characterized by its sleek, glossy finish and classic design, moves gracefully through a sequence of steps. The shoe's pointed toe and smooth surface reflect light subtly, emphasizing its refined craftsmanship. As it progresses, the shoe maintains a steady, rhythmic motion, suggesting a confident stride. The heel, slightly elevated, provides a sense of balance and stability, while the sole appears sturdy, designed for comfort and support. Throughout the sequence, the shoe's movement is fluid and purposeful, indicative of a person walking with intent and poise. The shoe's interaction with the ground is seamless, showcasing its flexibility and durability. The overall impression is one of elegance and sophistication, with the shoe serving as both a functional and stylish accessory.

(c) DAM is able to accurately keep track of small objects across the dynamic scene.

Input Video

DAM

A stone sculpture of a standing figure, draped in a flowing robe, is depicted in a serene and poised stance. The figure's right arm is bent at the elbow, with the hand gently resting on the chest, while the left arm is slightly extended outward, suggesting a gesture of contemplation or prayer. The head is adorned with a simple, rounded headpiece, and the facial features are finely detailed, conveying a sense of calm and introspection. The robe, intricately carved, wraps around the body, creating soft folds and drapes that add to the sculpture's lifelike quality. The figure stands firmly on a base, exuding a sense of stability and grace. Throughout the sequence, the sculpture remains stationary, embodying a timeless and tranquil presence.

(d) DAM correctly identifies stationary objects in the video, offering detailed descriptions for the appearance of the specified object.

Figure A.8: **Additional results from DAM on detailed localized video captioning (Part 1)**. Our model is able to accurately describe small objects in complex scenes that involve large object motion and camera motion. Our model also correctly identifies stationary objects captured by a non-stationary camera by saying they are stationary. Videos visualized in this figure are from SA-V [62] dataset.

labeling, resulting in 774k annotations across 593k images, as detailed in Tab. A.5. Due to filtering, the final number of instances and images is lower than the original 10% subset of SA-1B. We do not use the

Input Video

DAM

An elephant, characterized by its large, rounded ears and long, curved tusks, moves gracefully through the sequence. Its skin appears rough and textured, with a grayish-brown hue that suggests a natural, earthy tone. The elephant's trunk, a versatile and muscular appendage, is prominently visible, swaying gently as it walks. Its eyes are small and partially obscured by the folds of its skin, giving it a calm and serene expression. The elephant's legs are sturdy and columnar, supporting its massive body with ease. As it progresses through the sequence, the elephant maintains a steady and deliberate pace, its movements fluid and unhurried. The tusks, slightly curved and pointed, catch the light subtly, adding a touch of elegance to its robust form. Overall, the elephant exudes a sense of majesty and tranquility, moving with a quiet confidence that is both powerful and graceful.

(a) DAM accurately describes fine-grained details for the masked region.

Input Video

DAM

A white SUV is prominently featured, showcasing a sleek and modern design. The vehicle's body is smooth and aerodynamic, with a slightly elevated stance typical of SUVs. The front of the car is characterized by a distinctive grille and sharp, angular headlights that give it a bold and assertive look. As the sequence progresses, the SUV moves steadily forward, maintaining a consistent trajectory. The wheels, with their intricate alloy design, rotate smoothly, indicating a well-maintained vehicle. The side mirrors are positioned to provide optimal visibility, and the windows are tinted, adding a touch of sophistication. The SUV's movement is fluid, suggesting a well-engineered suspension system that handles the road with ease. Throughout the sequence, the car remains the focal point, exuding a sense of reliability and style.

(b) DAM is able to handle objects that are heavily occluded in some frames.

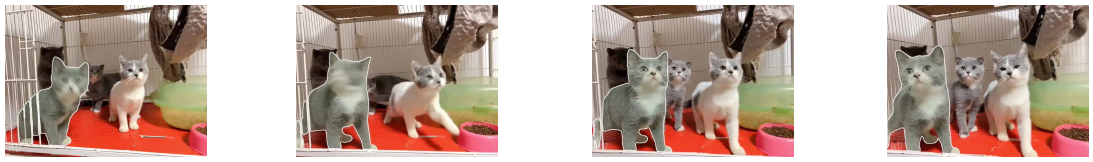
Input Video

DAM

A person wearing a blue jacket with white text on the back is captured in a sequence of dynamic movements. Initially, the individual is seen in a slightly bent posture, suggesting a moment of balance or readiness. As the sequence progresses, the person appears to be in motion, possibly walking or running, with their arms extended forward, indicating a sense of urgency or direction. The jacket, with its distinct color and text, remains a focal point throughout the sequence, emphasizing the person's presence and movement. The person's head is slightly tilted, and their posture suggests a forward-leaning stance, consistent with active movement. The sequence concludes with the person seemingly in mid-stride, capturing a sense of continuous motion and purpose.

(c) DAM is able to output detailed descriptions for objects despite occlusion due to object motion and human articulation.

Input Video



DAM

A cat with a sleek, short-haired coat of light gray and white fur is prominently featured. Its ears are perked up, and its large, expressive eyes are wide open, giving it an alert and curious demeanor. The cat's body is slightly turned, suggesting a poised and attentive stance. As the sequence progresses, the cat's head tilts slightly upward, maintaining its focused gaze. Its whiskers are long and prominent, adding to its inquisitive expression. The cat's posture remains mostly upright, with subtle shifts in weight that suggest a readiness to move or react. Throughout the sequence, the cat's overall demeanor is one of calm vigilance, as if it is observing something of interest with keen attention.

(d) DAM correctly identifies slight motion in the video, showing its capability for detailed motion analysis.

Figure A.9: **Additional results from DAM on detailed localized video captioning (Part 2)**. Our model is able to accurately describe objects that are partially occluded and is able to perceive and describe slight motion. Videos visualized in this figure are from MOSE [23] dataset.

masks provided with SA-1B, as they contain a large number of masks for parts. Instead, we employ the open-vocabulary detector OWL-ViT v2 [51, 52] to detect objects in the images, and then use SAM [62] to generate masks for the detected instances. Finally, we use SigLIP [97] to evaluate the image-text similarity,

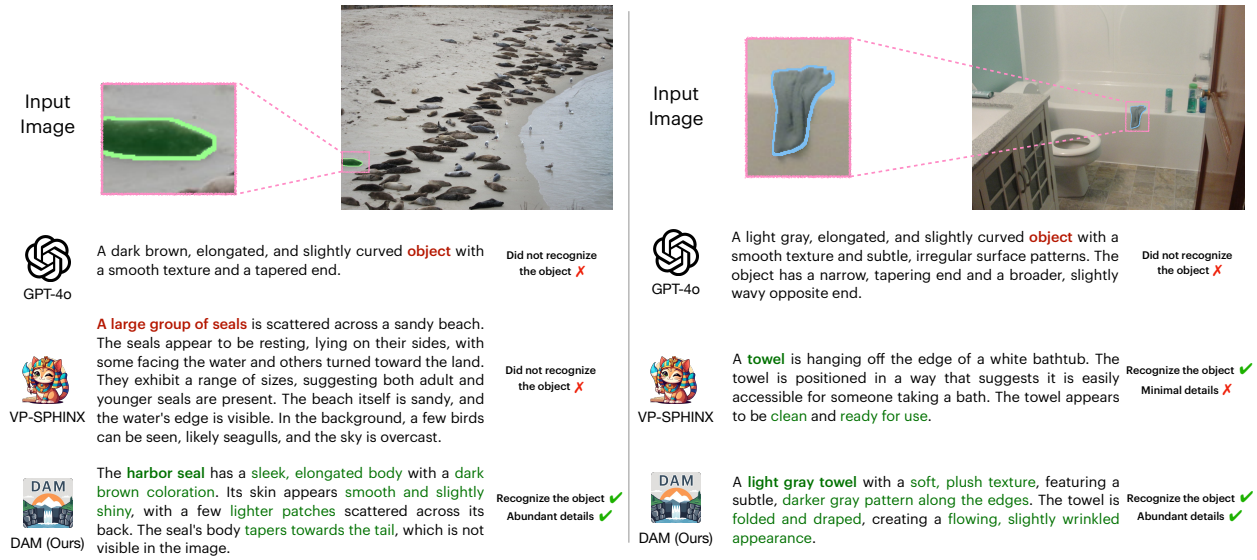


Figure A.10: **Qualitative comparisons demonstrate the superior localized image understanding capabilities of our model compared to GPT-4o [54] and VP-SPHINX [45], our strongest open-weight baseline**. GPT-4o struggles to recognize objects in masked regions accurately, offering only vague descriptions. In the left image, VP-SPHINX incorrectly describes a group of seals when the masked region contains only one seal. In the right image, VP-SPHINX identifies the towel but provides minimal detail, missing key attributes like its color. In contrast, our model delivers precise, detailed descriptions and captures the seal's sleek elongated body, dark brown coloration with lighter patches, and the towel's light gray color, wrinkled texture, and darker edge pattern. This superior performance stems from our model's architecture that effectively fuses object-specific details with broader contextual understanding.

taking the region as an image.

To ensure data quality, we apply extensive filtering (*i.e.*, rejection sampling) based on confidence scores from OWL-ViT v2, SAM, and SigLIP image-text similarity. We also ensure we have at most two instances per image, and for images with two instances, these two instances have to be from different classes. The object category names produced by OWL-ViT v2 are then put into a variant of our Describe Anything model, which is trained on data from stage 1 and optimized for self-labeling. This variant generates descriptions with a 50% probability of incorporating class names during training, as during self-labeling we have a class name as a part of each input. The object category proposals used by OWL-ViT v2 are generated by VILA 1.5 [44].

Detailed localized video captioning. We annotated 94k regions across 37k videos from SA-V dataset [62] for detailed localized video captioning, as detailed in Tab. A.6. Note that each region, also called masklet, indicates an instance across multiple frames in the video. In contrast to the use of SA-1B, where we did not use the masks that come with the dataset, we use the high-quality masklets that come with the videos. We found that many masklets cover parts of an instance, which is not necessarily helpful in describing the whole object as a common use case of our model. Therefore, we performed instance segmentation on the videos with ViTDet [41] + Cascade Mask R-CNN [12] trained by EVA-02 [26] and used voting to match the segmentation masks with the masklets. In this way, we filter out most of the masklets that are parts, since they likely do not correspond to instance masks. The matched masklets carry the class name from the matched instance segmentation mask, which is used in the annotation process to obtain a detailed localized caption for each masklet.

H.2. Model Training

We start from off-the-shelf VILA 1.5 [44] models that are publicly available on HuggingFace. For image-only training, we fine-tune VILA 1.5 3B model. For joint image-video training, we use VILA 1.5 8B model. We use SigLIP [97] vision encoder, following VILA 1.5. To prevent catastrophic forgetting and to maintain instruction following capabilities, we mix in ShareGPT-

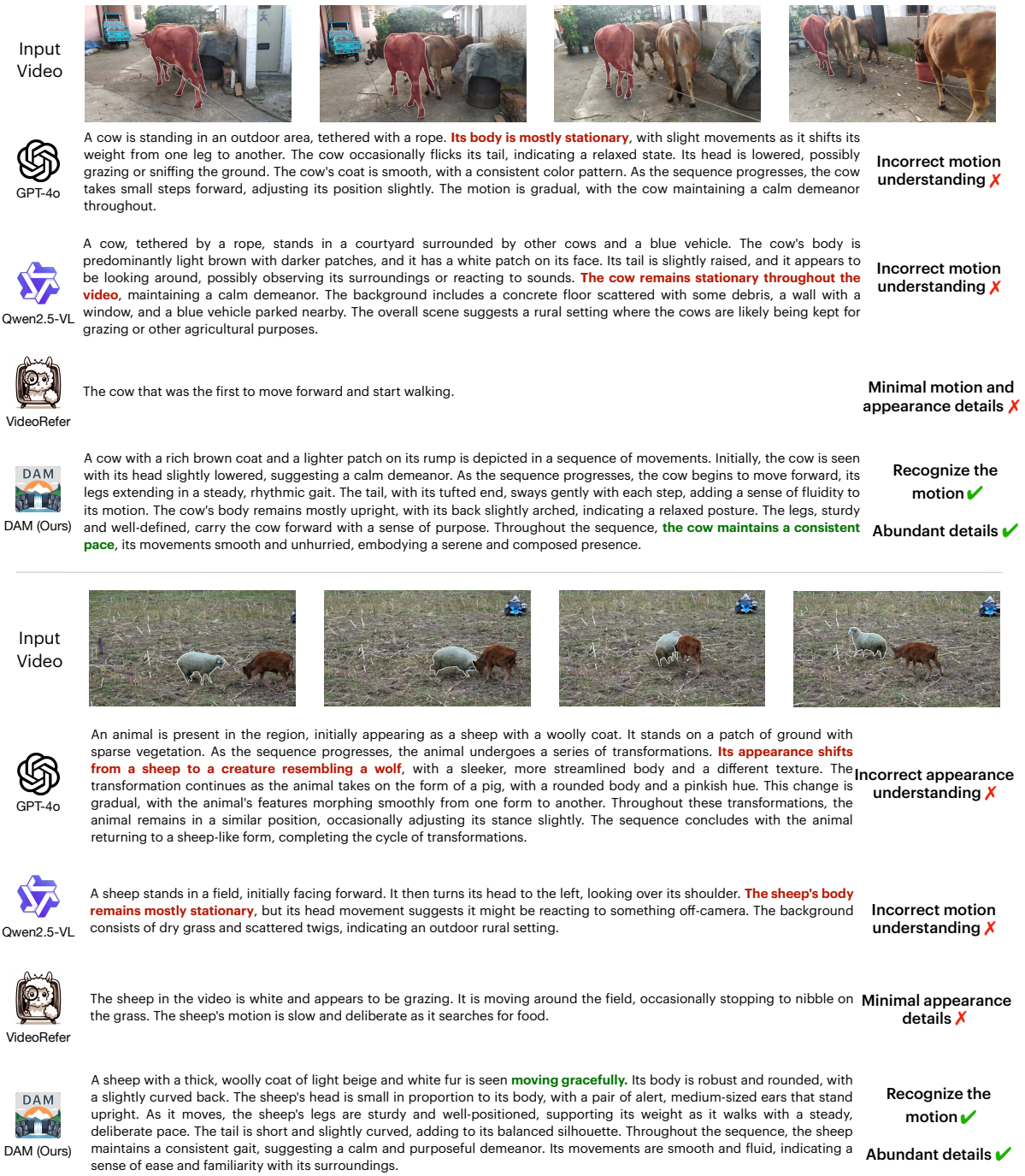


Figure A.11: **Our proposed DAM demonstrates superior localized video understanding compared to GPT-4o [54], QwenVL-2.5 [77], and VideoRefer [96]. Top figure:** DAM accurately captures the cow's forward movement with comprehensive details, whereas GPT-4o and QwenVL-2.5 mistakenly perceive the cow as stationary. Compared to VideoRefer, DAM provides richer descriptions of both motion and appearance. **Bottom figure:** DAM correctly recognizes the animal as a sheep and accurately describes its graceful movement, while GPT-4o erroneously identifies it as transforming into other animals, and QwenVL-2.5 incorrectly perceives that only the sheep's head is moving. VideoRefer provides limited appearance details, while DAM offers extensive, accurate descriptions. These cases highlight DAM's precise understanding of motion and appearance throughout video sequences.

```
You are responsible to write a very descriptive caption to describe the {{category}} in the provided SEGMENTED image. You may leverage the surrounding context of the SEGMENTED image provided in the CROPPED image.
You must not mention any background in the caption and only describe the {{category}} in the SEGMENTED image! The caption must ONLY contain sufficient details to reconstruct the same {{category}} in the SEGMENTED image but nothing else!
Here are some additional rules you need to follow when describing the {{category}} in the SEGMENTED image:
1. If there are multiple {{category}} in the CROPPED image, focus on the {{category}} in the SEGMENTED image.
2. If the {{category}} in the SEGMENTED image is occluded by other objects, only describe the visible part. DO NOT mention anything that is not directly related to the visible part of {{category}}, such as "A segment of", which part is invisible, etc. For objects with text written on it, describe the object instead of just outputting the text written on it.
Here is the SEGMENTED image that needs caption:
```

Table A.7: Our prompt for data annotation in stage 1.

4V [16] with our localized image/video captioning dataset collected with our proposed data pipeline. Following the VILA 1.5 training and inference recipe, we treat videos as 8 images concatenated in the sequence.

We closely follow VILA 1.5's recipe of the supervised fine-tuning stage and train all modules, including the vision backbone, the projector, and the LLM. We fine-tune the model for 1 epoch. For the 3B model, we use a batch size of 2048 with a learning rate of 1e-4 on 8 Nvidia A100 GPUs. For the 8B model, we use a batch size of 2048 with a learning rate of 1e-5 on 32 Nvidia A100 GPUs. Both models take less than a day to train. We use a cosine scheduler with a warmup ratio of 0.03. No weight decay is used. For training our model that takes in a class name for self-labeling, we randomly put the class name in the prompt with 50% probability. For models without prompt augmentation, which is detailed below, we simply use the prompt "`Describe the masked region in detail.`" Following VILA, we always put image tokens in front of the textual tokens. As for the setting for the focal crop, we extend the crop by $1\times$ the width towards left and right, and $1\times$ the height towards top and bottom, unless we hit the boundaries of the image, in which case we take the boundaries, *i.e.* $\alpha = 3$ and the total area of the crop is enlarged up to $9\times$. If either the height or width is less than 48 pixels, we take 48 pixels for that direction to encode more context for very small regions, since the small regions themselves do not have much useful information.

Prompt Augmentation. We trained a variant of our model with prompt augmentation to enhance generalization capabilities beyond detailed localized captioning, as analyzed in App. G. For these models, during training, we randomly select one of 15 prompts from a predefined set. These prompts may or may not include a `{prompt_suffix}`. The default prompt suffix is *in detail*. However, we introduce variability by conditioning the prompt on the number of words or sentences in the target caption.

Specifically, with a 20% probability, we condition the prompt on the number of sentences, using suffixes like *in one sentence* or *in [number of sentences] sentences* (e.g., *in 2 sentences*). If the caption contains only one sentence, we use phrases like *in a sentence* or *in one sentence*.

With another 20% probability, we condition the prompt on the number of words in the target caption. For captions with a small word count, we use exact numbers (e.g., *in 3 words*). For longer captions (up to 200 words), we may round the word count to the nearest ten and use phrases like *in about 50 words* or *in around 50 words*. If the caption exceeds 200 words, we use the suffix *in more than 200 words*.

The list of prompts that include a `{prompt_suffix}` is as follows:

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}`?

Additionally, we have prompts that inherently request detailed descriptions without requiring a suffix:

1. Give a detailed description of the masked region.
2. Provide a thorough description of the masked region.
3. Can you explain the details of the masked area?
4. Give a detailed account of the masked region.
5. Describe the masked area comprehensively.
6. Provide an in-depth description of the masked region.
7. Explain the specifics of the masked area.
8. Can you provide a thorough explanation of the masked region?
9. What are the details of the masked area?
10. Provide a comprehensive description of the masked area.
11. What specific details can you provide about the masked region?
12. Can you give an in-depth account of the masked section?
13. What are the main characteristics of the masked region?
14. Give a thorough description of the masked area's details.
15. Provide detailed information about the masked area.

For prompts without a suffix, we do not condition the generation on the number of words or sentences.

During training, we select prompts based on the `prompt_suffix`:

- If the `prompt_suffix` is *in detail* (the default option), we may choose from either set of prompts.
- If the `prompt_suffix` specifies word or sentence counts, we select only from prompts that include `{prompt_suffix}`.

This approach introduces variability in the prompts, encouraging the model to generate responses with controls from the prompts in mind, thereby enhancing its generalization and instruction-following capabilities.

H.3. Inference Setting

Unless otherwise mentioned, our prompt for obtaining detailed localized image descriptions at inference time is the following:

```
Describe the masked region in detail.
```

Our prompt for obtaining detailed localized video descriptions at inference time is the following:

```
Given the video in the form of a sequence
of frames above, describe the object in
the masked region in the video in detail.
Focus on appearance, motion, and actions.
If the motion involves multiple stages or
steps, break down each stage and describe
the movements or changes sequentially.
Ensure each phase of motion is described
clearly, highlighting transitions between
actions.
```

For Co3Dv2 [64] sequences that we treat as videos, we use the following prompt:

```
Describe the masked region in the video
in detail.  The video consists of multiple
views of a stationary object.  Focus on the
appearance of the object without mentioning
any motion or actions.
```

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 3, 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 2, 3, 8

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv:2308.01390*, 2023. 3

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv:2309.16609*, 2023. 3

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005. 2, 3, 8

[6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Fuyu-8b: A multimodal architecture for ai agents, 2023. 3

[7] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv:1911.09785*, 2019. 2, 6

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019. 2, 6

[9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv:2407.07726*, 2024. 3

[10] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv:2210.09461*, 2022. 4

[11] H Caesar, J Uijlings, and V Ferrari. Coco-stuff: Thing and stuff classes in context. arxiv. *arXiv:1612.03716*, 2016. 11

[12] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *TPAMI*, 2019. 14

[13] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. *arXiv:2310.12971*, 2023. 8, 3

[14] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv:2406.16866*, 2024. 7, 8, 3

[15] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 1, 3, 7, 8

[16] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023. 16

[17] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024. 8, 9, 3

[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015. 1

[19] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. On scaling up a multilingual vision and language model. In *CVPR*, 2024. 3

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv:2312.14238*, 2023. 3, 8, 9, 10

[21] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv:2412.05271*, 2024. 8, 10

[22] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv:2305.06500*, 2023. 3

[23] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *CVPR*, 2023. 13

[24] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 5

[25] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 3, 8, 6, 10

[26] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024. 14

[27] Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. Dataseg: Taming a universal multi-dataset multi-task segmentation model. *NeurIPS*, 2024. 6

[28] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *CVPR*, 2024. 2, 3, 7, 8, 5, 6

[29] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 7, 2, 9, 11

[30] Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. In *CVPR*, 2024. 1, 3, 6, 8

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 3

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4, 1

[33] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 11

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2, 3

[35] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 11

[36] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *NeurIPS*, 2024. 3

[37] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 2

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3, 5

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

[40] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan

Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. 2022 ieee. In *CVPR*, pages 10955–10965, 2021. 1

[41] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 14

[42] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 1

[43] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2, 3, 8

[44] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *CVPR*, 2024. 3, 6, 8, 5, 10, 14

[45] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *ICLR*, 2025. 1, 2, 3, 7, 8, 9, 10, 14

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 8, 1, 10

[47] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 3, 7, 8, 1, 10

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 1

[50] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2, 3

[51] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 13

[52] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 2024. 13

[53] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 11

[54] OpenAI. Gpt-4o system card, 2024. 1, 2, 6, 8, 9, 3, 10, 14, 15

[55] OpenAI. Learning to reason with llms, 2024. 2, 6, 8, 10

[56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 2, 3, 8

[57] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 3

[58] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 7, 8, 3

[59] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. *NeurIPS*, 37:114321–114347, 2024. 3, 8, 9

[60] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023. 7, 2, 11

[61] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 3, 7

[62] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 4, 6, 1, 11, 12, 13, 14

[63] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[64] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 9, 17

[65] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. In *CVPR*, 2022. 4

[66] Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *CVPR*, 2023. 4

[67] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3, 6

[68] Manish Kumar Singh, Rajeev Yasarla, Hong Cai, Mingu Lee, and Fatih Porikli. Tosa: Token selective attention for efficient vision transformers. *arXiv:2406.08816*, 2024. 4

[69] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020. 2, 6

[70] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, 2024. 3

[71] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 8, 3

[72] Claude Team. Claude 3.5 sonnet, 2024. 6

[73] Claude Team. Claude 3.7 sonnet and claude code, 2025. 8, 10

[74] Gemini Team. Gemini 2.5: Our most intelligent ai model. 2025. 8, 10

[75] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 8, 6, 10

[76] Gemini Team, M Reid, N Savinov, D Teplyashin, Lepikhin Dmitry, T Lillicrap, JB Alayrac, R Soricut, A Lazaridou, O Firat, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024. 6

[77] Qwen Team. Qwen2.5-vl, 2025. 8, 10, 15

[78] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 5

[79] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 3, 8

[80] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *ECCV*, pages 166–185. Springer, 2024. 9

[81] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. 3, 8, 9, 10

[82] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls. *arXiv:2305.02677*, 2023. 1, 3

[83] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv:2308.01907*, 2023. 3

[84] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv:2411.10442*, 2024. 8, 10

[85] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv:2212.00280*, 2022. 1

[86] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv:2312.14135*, 2023. 4

[87] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv:2409.04429*, 2024. 3

[88] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 2

[89] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 6

[90] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv:2107.00641*, 2021. 4

[91] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv:2310.11441*, 2023. 3, 1, 8, 10

[92] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv:2309.17421*, 2023. 3, 1

[93] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024. 1, 2, 3, 7, 8, 9, 5

[94] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *ECCV*, pages 425–443. Springer, 2024. 3, 8

[95] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024. 1, 2, 3, 7, 8, 9

[96] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *arXiv:2501.00599*, 2024. 2, 3, 8, 9, 4, 10, 15

[97] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 5, 13, 14

[98] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. In *COLM*, 2024. 3

[99] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv:2307.03601*, 2023. 1, 3, 7, 8

[100] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv:2406.19389*, 2024. 1, 3, 8, 6

[101] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. Controlcap: Controllable region-level captioning. pages 21–38, 2024. 1, 3, 8

[102] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1

[103] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 3