

## 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 spe-

cific 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.
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.

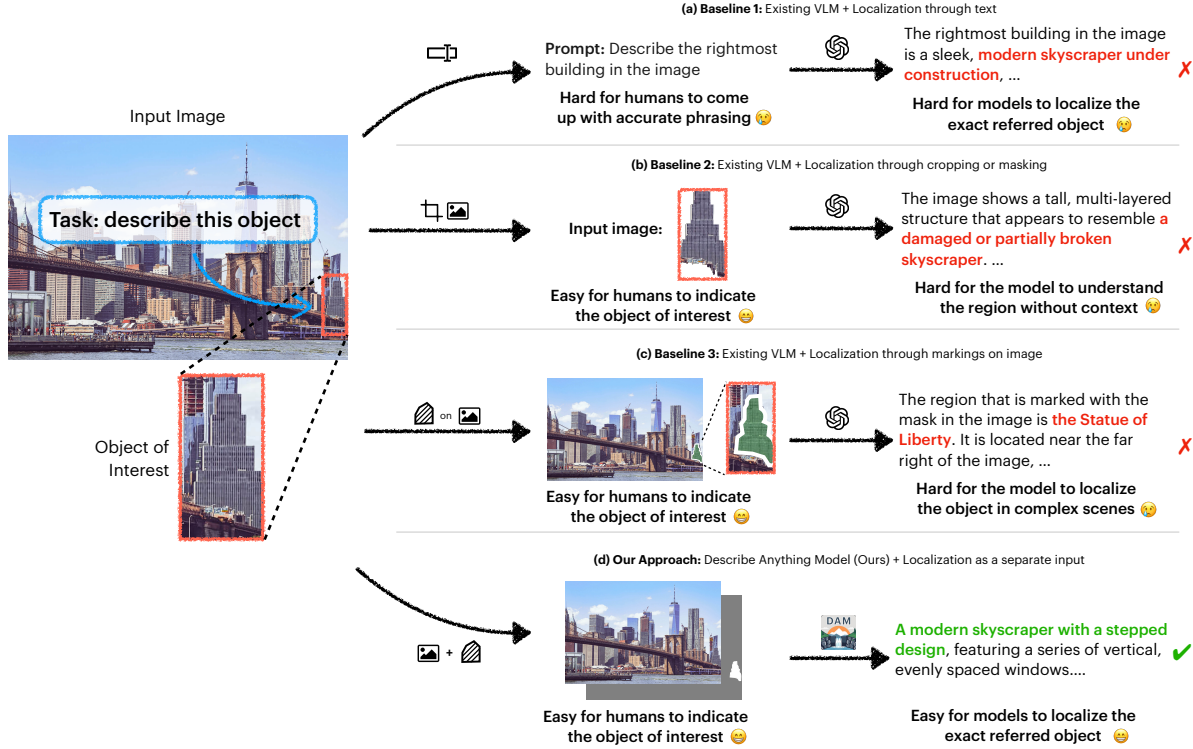


Figure A.1. **Existing Vision-Language Models (VLMs) do not perform well in generating localized descriptions.** (a) to (c) demonstrate several ways to prompt existing VLMs, but none achieves satisfactory performance, leading to the need for a new method that is capable of providing detailed and localized 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.

- 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.

- 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.
- 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 Appendix D. Results are in Tab. 5.



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.

#### 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.

## 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 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 Secs. 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.


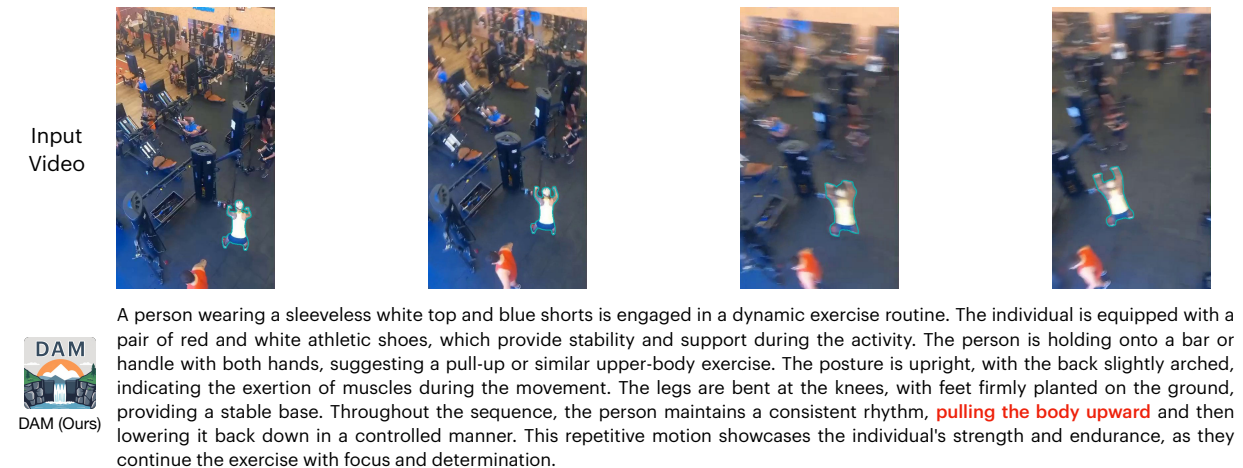
Input Video	
DAM Output	A man with short, dark hair and a slight stubble is wearing a denim shirt with rolled-up sleeves. He begins by standing <b>with his hands clasped together</b> in front of him, exuding a calm and composed demeanor. As the sequence progresses, he <b>raises his hands slightly</b> , maintaining a gentle and expressive gesture. His facial expression shifts subtly, suggesting he is engaged in <b>conversation or explanation</b> . The man's movements are smooth and deliberate, as he occasionally <b>opens his hands and gestures outward</b> , indicating emphasis or clarification. Throughout the sequence, his posture remains upright and attentive, conveying a sense of engagement and focus.
Reference Caption	A man with short black hair is standing on the left, wearing a black jacket, as if reporting news
GPT Evaluation on Hallucination Detection (HD) Dimension	Hallucination Detection: 1 Explanation: The predicted answer includes several imaginative elements, such as <b>gestures and expressions</b> , that are not mentioned in the correct answer, indicating hallucinations in the description.

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.



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



(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.



### 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, 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 Appendix 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 deduplication 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

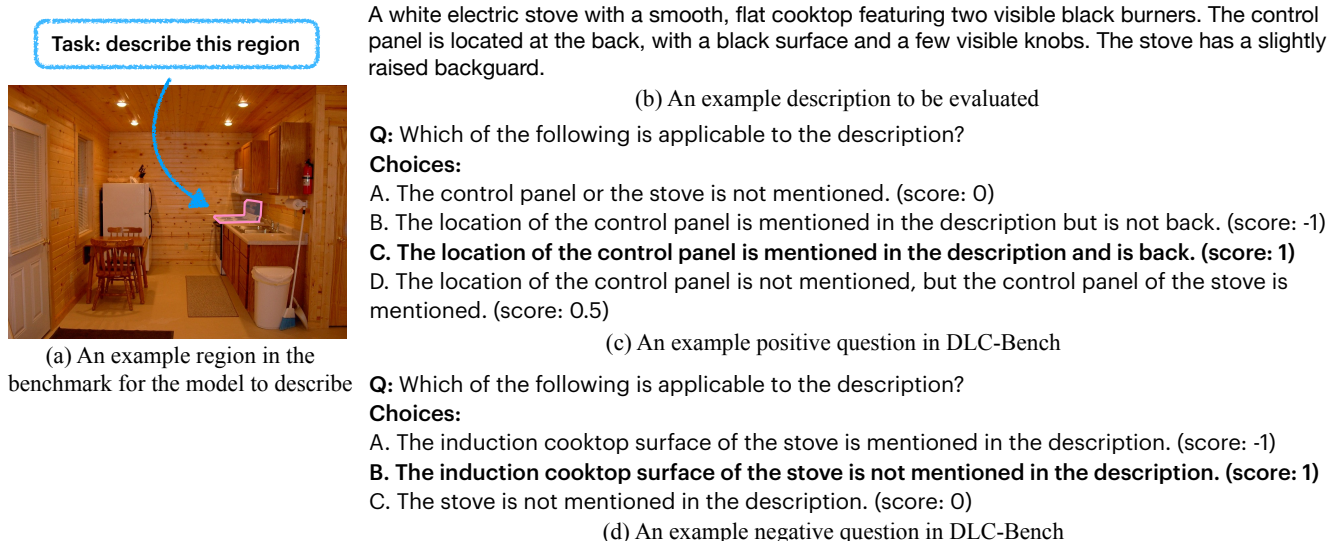


Figure A.5. **An example from DLC-Bench for detailed localized captioning.** (a) The process begins by prompting a model to describe a specified region within the image. The resulting description is then evaluated using a text-only LLM as a judge that rates each response by answering positive and negative 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.

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 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 Appendix 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 Appendix 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.

## 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 local-

	VP-SPHINX Arch	Our Arch
Avg (%)	50.2	<b>63.8</b>

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	<b>52.3</b>	<b>82.2</b>	<b>67.3</b>
Yes	51.3	<b>82.2</b>	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	<b>52.4</b>	<b>85.4</b>	<b>68.9</b>

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.

ized 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.

## 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.



Method	#Params	Pos (%)	Neg (%)	Avg (%)
<i>API-only General VLMs:</i>				
GPT-4o (SOM) [54]	-	5.0	29.2	17.1
o1 (SOM) [55] <sup>†</sup>	-	0.8	28.0	14.4
Claude 3.7 Sonnet (SOM) [73] <sup>†</sup>	-	0.5	40.2	20.4
Gemini 2.5 Pro (SOM) [74, 75] <sup>†</sup>	-	13.2	65.0	39.1
<i>Open-source General VLMs:</i>				
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
<b>DAM (Ours)</b>	<b>3B</b>	<b>52.3</b>	<b>82.2</b>	<b>67.3</b>

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.

**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 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

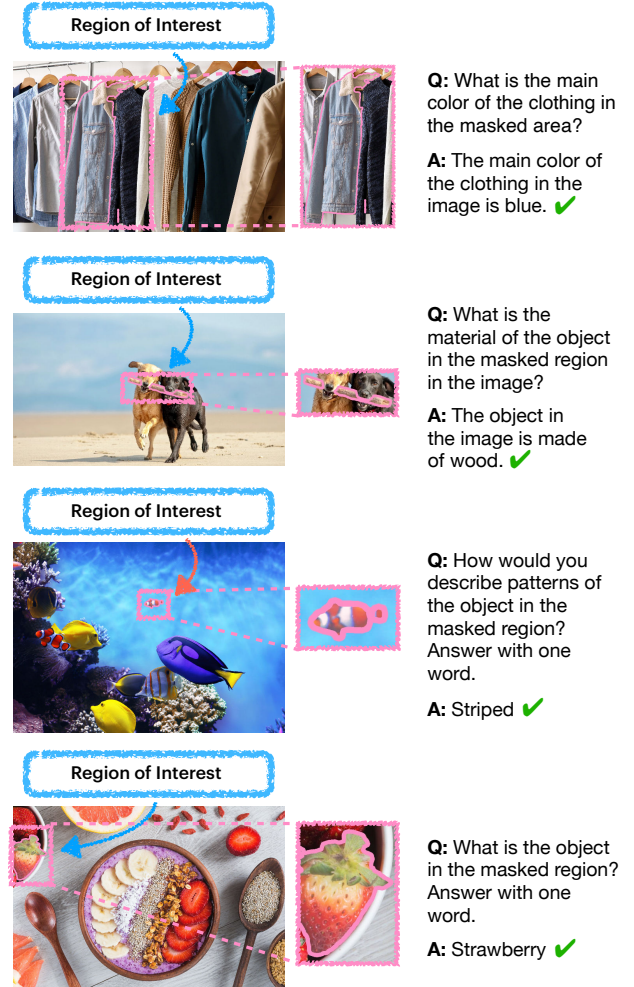


Figure A.6. **Emerging zero-shot QA capabilities.** DAM could answer questions about regions in an image, showcasing capabilities such as object recognition and property identification.

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-labeling,

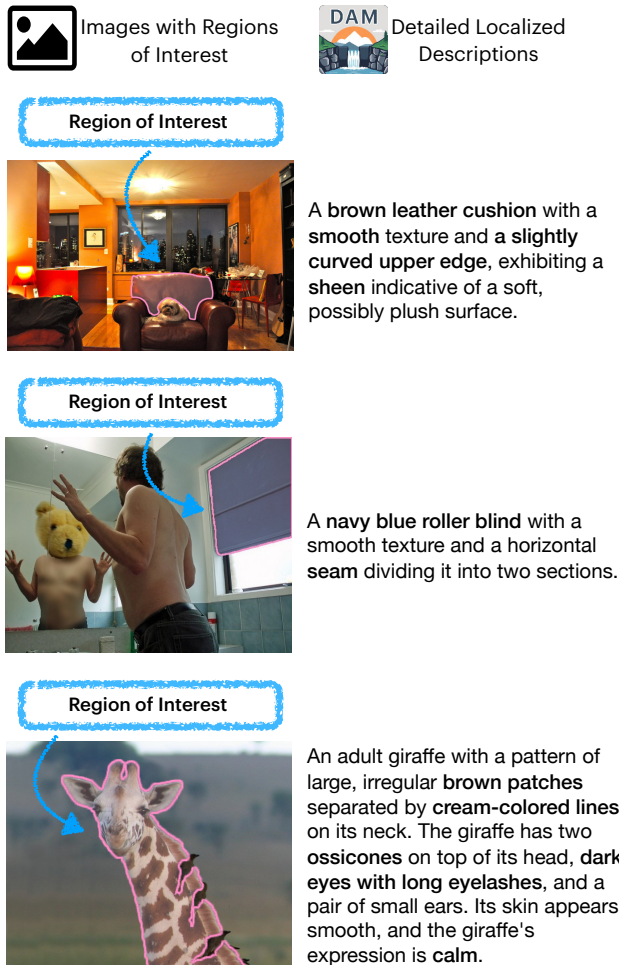


Figure A.7. **Additional results from LVIS [29] demonstrating DAM’s detailed localized image captioning capabilities.** Our model exhibits robust region understanding and localization across diverse scenarios. It produces precise descriptions of objects within masked regions while successfully identifying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

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 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, 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

Dataset	# Images	# Regions
<i>Stage 1:</i>		
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
<i>Stage 2:</i>		
SA-1B (10%)	592,822	774,309
<b>Total</b>	<b>819,035</b>	<b>1,458,203</b>

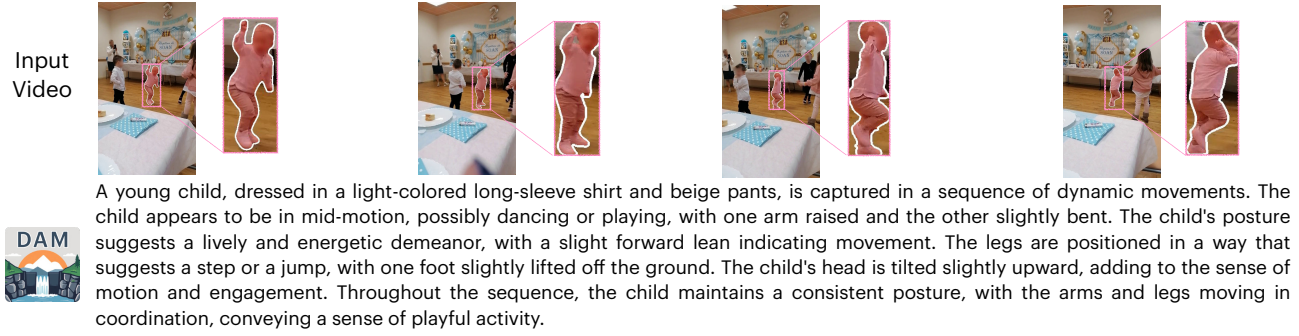
Table A.5. **Dataset statistics across stages with total images and regions for training detailed localized image captioning.** In stage 1, we annotated 684k regions across 226k images from existing instance and semantic segmentation datasets. In stage 2, we perform SSL on 10% of SA-1B images without 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

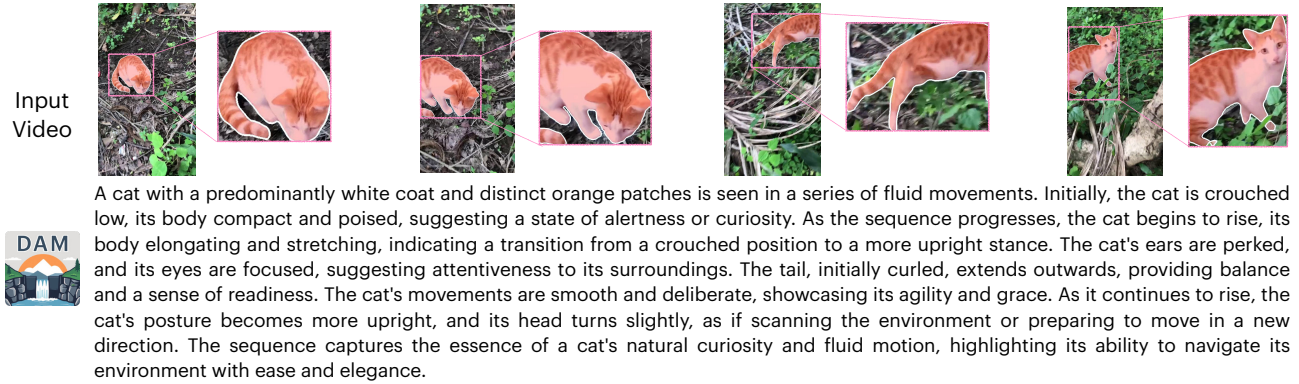
Table A.6. **Dataset statistics across stages with total videos and regions for training detailed localized video captioning.** We label 94k regions across 37k videos from SA-V dataset [62] for detailed localized video captioning. Note that each region indicates an instance across multiple frames in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

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].

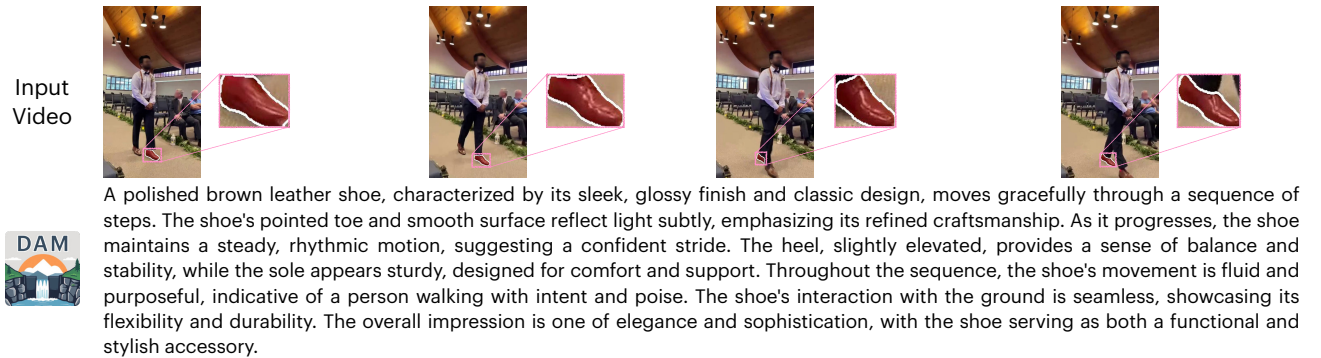
**Detailed localized video captioning.** We annotated 94k regions across 37k videos from SA-V dataset [62] for detailed localized video captioning, as detailed in Tab. A.6. Note that each region, also called masklet, indicates an instance across multiple frames in the video. In contrast to the use of 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



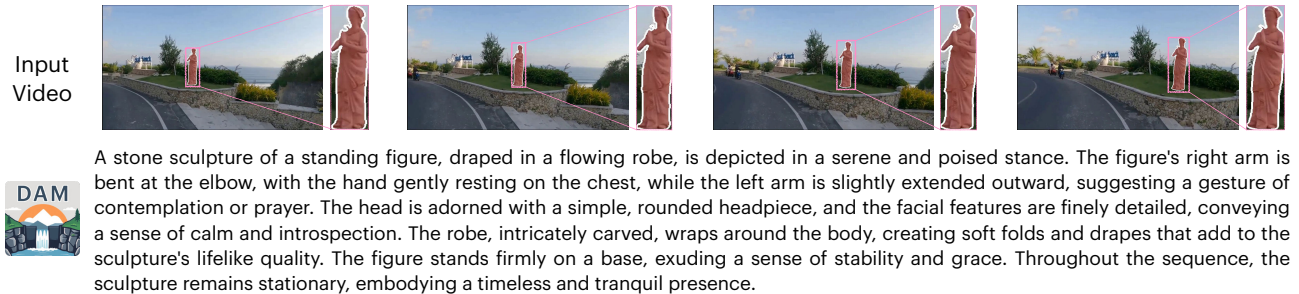
(a) DAM accurately describes fine-grained details for the masked region.



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



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



(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.



Input  
Video



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.

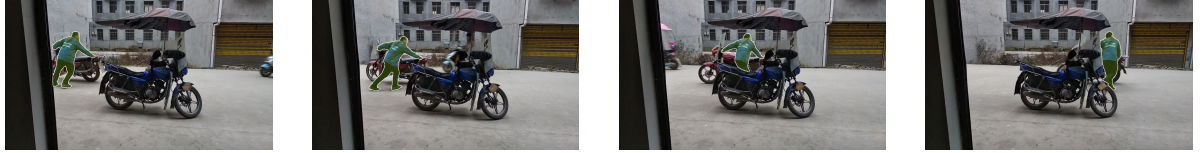
Input  
Video



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.

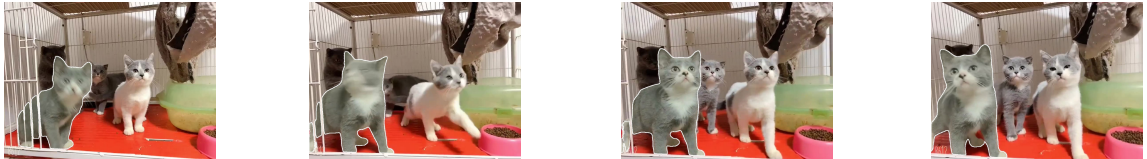
Input  
Video



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



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.



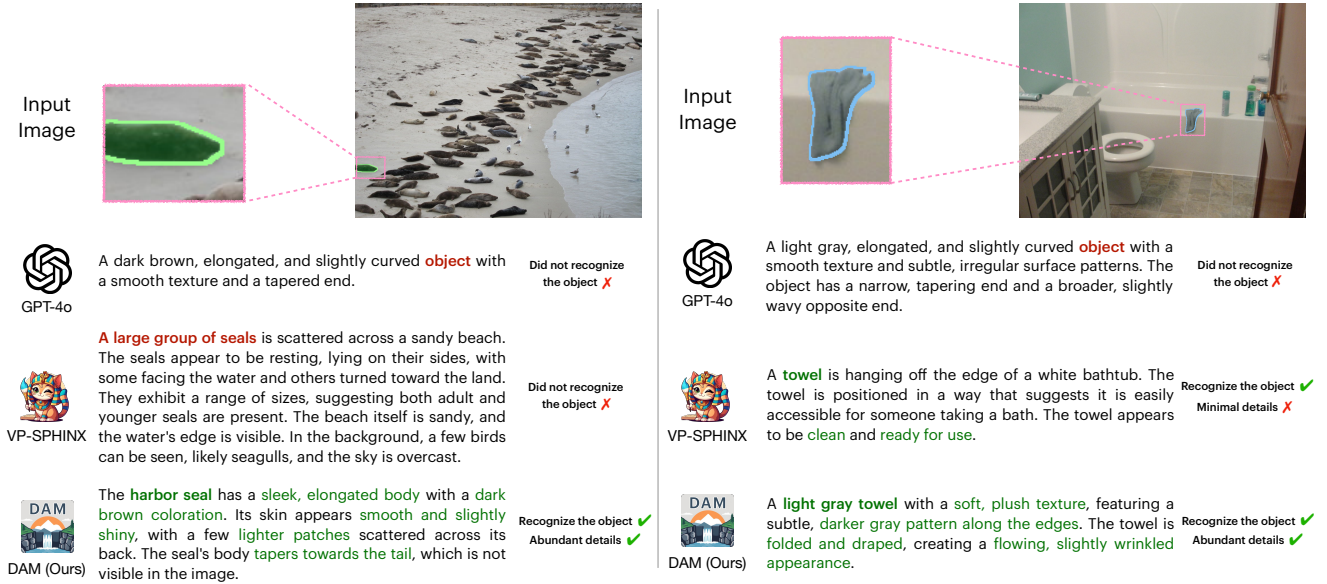


Figure A.10. **Qualitative comparisons demonstrate the superior localized image understanding capabilities of our model compared to GPT-4o [54] and VP-SPHINX [45], our strongest open-weight baseline.** GPT-4o struggles to recognize objects in masked regions accurately, offering only vague descriptions. 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.

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-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 Appendix G. For these models, during training, we randomly select one of 15 prompts from a pre-

- 
- 1 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.
  - 2 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!
  - 3 Here are some additional rules you need to follow when describing the {{category}} in the SEGMENTED image:
  - 4 1. If there are multiple {{category}} in the CROPPED image, focus on the {{category}} in the SEGMENTED image.
  - 5 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.
  - 6 Here is the SEGMENTED image that needs caption:
- 

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

defined 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.











Input Video		
 GPT-4o	A cow is standing in an outdoor area, tethered with a rope. <b>Its body is mostly stationary</b> , with slight movements as it shifts its weight from one leg to another. The cow occasionally flicks its tail, indicating a relaxed state. Its head is lowered, possibly grazing or sniffing the ground. The cow's coat is smooth, with a consistent color pattern. As the sequence progresses, the cow takes small steps forward, adjusting its position slightly. The motion is gradual, with the cow maintaining a calm demeanor throughout.	Incorrect motion understanding <b>X</b>
 Qwen2.5-VL	A cow, tethered by a rope, stands in a courtyard surrounded by other cows and a blue vehicle. The cow's body is predominantly light brown with darker patches, and it has a white patch on its face. Its tail is slightly raised, and it appears to be looking around, possibly observing its surroundings or reacting to sounds. <b>The cow remains stationary throughout the video</b> , maintaining a calm demeanor. The background includes a concrete floor scattered with some debris, a wall with a window, and a blue vehicle parked nearby. The overall scene suggests a rural setting where the cows are likely being kept for grazing or other agricultural purposes.	Incorrect motion understanding <b>X</b>
 VideoRefer	The cow that was the first to move forward and start walking.	Minimal motion and appearance details <b>X</b>
 DAM (Ours)	A cow with a rich brown coat and a lighter patch on its rump is depicted in a sequence of movements. Initially, the cow is seen with its head slightly lowered, suggesting a calm demeanor. As the sequence progresses, the cow begins to move forward, its legs extending in a steady, rhythmic gait. The tail, with its tufted end, sways gently with each step, adding a sense of fluidity to its motion. The cow's body remains mostly upright, with its back slightly arched, indicating a relaxed posture. The legs, sturdy and well-defined, carry the cow forward with a sense of purpose. Throughout the sequence, <b>the cow maintains a consistent pace</b> , its movements smooth and unhurried, embodying a serene and composed presence.	Recognize the motion <b>✓</b> Abundant details <b>✓</b>
Input Video		
 GPT-4o	An animal is present in the region, initially appearing as a sheep with a woolly coat. It stands on a patch of ground with sparse vegetation. As the sequence progresses, the animal undergoes a series of transformations. <b>Its appearance shifts from a sheep to a creature resembling a wolf</b> , with a sleeker, more streamlined body and a different texture. The transformation continues as the animal takes on the form of a pig, with a rounded body and a pinkish hue. This change is gradual, with the animal's features morphing smoothly from one form to another. Throughout these transformations, the animal remains in a similar position, occasionally adjusting its stance slightly. The sequence concludes with the animal returning to a sheep-like form, completing the cycle of transformations.	Incorrect appearance understanding <b>X</b>
 Qwen2.5-VL	A sheep stands in a field, initially facing forward. It then turns its head to the left, looking over its shoulder. <b>The sheep's body remains mostly stationary</b> , but its head movement suggests it might be reacting to something off-camera. The background consists of dry grass and scattered twigs, indicating an outdoor rural setting.	Incorrect motion understanding <b>X</b>
 VideoRefer	The sheep in the video is white and appears to be grazing. It is moving around the field, occasionally stopping to nibble on the grass. The sheep's motion is slow and deliberate as it searches for food.	Minimal appearance details <b>X</b>
 DAM (Ours)	A sheep with a thick, woolly coat of light beige and white fur is seen <b>moving gracefully</b> . Its body is robust and rounded, with a slightly curved back. The sheep's head is small in proportion to its body, with a pair of alert, medium-sized ears that stand upright. As it moves, the sheep's legs are sturdy and well-positioned, supporting its weight as it walks with a steady, deliberate pace. The tail is short and slightly curved, adding to its balanced silhouette. Throughout the sequence, the sheep maintains a consistent gait, suggesting a calm and purposeful demeanor. Its movements are smooth and fluid, indicating a sense of ease and familiarity with its surroundings.	Recognize the motion <b>✓</b> Abundant details <b>✓</b>

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.