Seeing What Matters: A Training-Free Self-Guided Framework for Multimodal Detail Perception and Reasoning

Supplementary Material

8. Implementation Details

We use python 3.10.16, transformers 4.47.0, scikit-learn 1.6.1, torch 2.6.0, and torchvision 0.21.0. Our hardware environment consists of Intel(R) Xeon(R) Silver 4310 CPUs with 128 GB of RAM, and 4 NVIDIA RTX4090-24GB GPUs for all of our experiments. We use "short questions" during *scan-locate* stage and full prompt during *focus* inference. Specifically, for TextVQA, SLoFo first use the question in the validation annotation file to form a prompt without external OCR tokens, unlike the baseline LLaVA-v1.5 original settings [25]. Since our goal is to enable the model to select the most relevant visual regions during this stage, we only use images and questions while avoiding using OCR extracted token. As for the focus stage, we provide the OCR tokens for external guidance and help with token pruning. For a fair comparison, we reproduced ViCrop-LLaVA [40] and also provided these additional OCR information in its input prompt. And for multi-choice tasks such as V*, SLoFo only use the questions and images to form prompts, without use of options, during *scan-locate* stage.

9. Dataset Statistics

In this section, we present the details and statistics of the benchmarks used to evaluate SLoFo.

B.1 Detail-Sensitive Tasks

Following the categorization in [40], we consider TextVQA [32], POPE [21], DocVQA [29], and V* [36] as detail-sensitive tasks. To further investigate MLLMs' perception of fine-grained visual details under higher resolutions, we additionally classify these four tasks as **Small Detail Perception Tasks**, and introduce HR-Bench-4K and HR-Bench-8K [35] in parallel as **Tiny Detail Perception Tasks**.

1. **TextVQA** The validation set contains 5,000 questions across 3,166 images. For the hyper-parameter studies presented in Figure 3 and 4, we use a fixed subset of the validation set by randomly selecting 800 images and sampling one question per image. This same subset is used to evaluate DeepSeek-VL-7B. Since it doesn't comply with given instructions (see instructions in the Appendix C.) and instead of answering with a single word or phrase as instructed, it leans to output complete sentence or apologize if it "doesn't know" the answer. We have to evaluate DeepSeek-VL-7B answers manually, so we choose to test it with the same subset.
2. **POPE** (Polling-based Object Probing Evaluation) provides a general benchmark for evaluating visual hallucination in MLLMs. It tests the accuracy of object presence recognition through balanced queries containing both existing and non-existing objects (50% each). The benchmark includes three types of negative object sampling strategies: (1)*Random:* objects not present in the image are randomly selected; (2) *Popular:* objects not in the image but frequently occurring overall are selected; (3) *Adversarial:* objects that commonly co-occur with existing ones, but are not present in the image, are selected. POPE is built upon three image sources: MSCOCO, A-OKVQA, and GQA. For each dataset and each sampling strategy, 500 images are randomly selected, and 6 queries are generated per image—yielding 9,000 query-answer pairs per dataset and a total of 27,000 examples. In Table 1, we report the average accuracy across the three strategies on MSCOCO. Table 4 presents the full breakdown of Accuracy, Precision, Recall, and F1 score on all three datasets under all three settings, comparing the baseline, SLoFo, and SLoFo high-res.
3. **DocVQA** (Document Visual Question Answering) is an open-ended VQA benchmark for document images, aiming to drive research in document analysis and understanding. We evaluate on its validation set, which contains 1,286 images and 5,349 questions.
4. **V*** is a benchmark focused on fine-grained details in high-resolution images, designed to test the ability of models to reason over small objects or object features. It consists of 191 image-question pairs, including single-choice questions divided into two categories: *Direct attribution* targeting at attributes like color or pose; and *Spatial relationship* evaluating location-based reasoning between objects.
5. **HR-Bench** is designed to rigorously evaluate on ultra-high-resolution images. The 4K and 8K subsets each contain 800 image-question pairs. These examples pose significant challenges for detail reasoning and serve as extreme settings for evaluating visual focus ability.

9.1. General-Purpose Tasks

To evaluate the generalizability of our method beyond detail-sensitive scenarios, we conduct experiments on several widely-used general-purpose vision-language benchmarks: VQA, GQA, SEED, SQA, VizWiz, and MME.

1. **VQAv2** is a standard benchmark for evaluating im-

Figure 6. Success (upper 6) and failure (lower 3) examples of SLoFo on TextVQA.

age understanding and language reasoning. We use the VQAv2 test-standard set, which contains 447,793 image-question pairs covering a wide range of reasoning types, including yes/no, number, and other open-ended answers.

2. **GQA** (Graph Question Answering) is a large-scale dataset emphasizing compositional reasoning and scene understanding. We evaluate on the testdev_balanced set, consisting of 12,578 question-image pairs.
3. **SEED-Bench** is a recent benchmark designed to evaluate and MLLM capabilities in image and video reasoning. We use the image part of the benchmark including 14,233 input pairs.
4. **SQA** (ScienceQA) is a multimodal science question-answering benchmark covering elementary and middle school science questions. It includes both textual and visual questions, often requiring reading diagrams, charts, or textbook figures. We evaluate on multimodal questions in validation set.
5. **VizWiz** is a real-world VQA benchmark collected from blind users. The dataset includes images with noise, occlusion, or low quality, and questions often contain typos or colloquial expressions. We use the test set with 8,000 image-question pairs, which presents significant challenges for standard models.
6. **MME**, the first comprehensive MLLM evaluation benchmark, tests MLLMs across multiple fine-grained capabilities. It includes question categories such as object counting, OCR recognition, spatial understanding, commonsense, etc. We follow standard evaluation and report the perception and cognition scores across all subtasks.

10. Prompt Settings

We provide details about the prompt format we use for zero-shot inference. For tasks involving OCR tokens, choices or other external information, we only use the question itself (i.e. short_question) during the *Scan-Locate* stage. And these necessary information (i.e. external_info) are concatenated in the prompt during *Focus* inference.

SLoFo *Scan-Locate* stage

```
<image> USER:{short_question}
Answer the question using a single
word or phrase.    ASSISTANT:
```

SLoFo *Focus* stage

```
<image> <sub-image>
USER:{question} {external_info}
Answer the question using a single
word or phrase.    ASSISTANT:
```

11. Additional Examples

In Figure 6 and 7, we present more examples on TextVQA and V*, including success scenarios, where the LLaVA baseline answers incorrectly and our SLoFo framework helps correct it, as well as failure scenarios, where the baseline initially answers the question correctly but SLoFo generates incorrect answers.

12. Additional Inference Overhead Analysis

To provide a theoretical perspective, the per-layer MACs of a Transformer decoder follows:

$$T(N) = 4Nd^2 + 2N^2d + 3Ndd_m, \tag{5}$$

where the $2N^2d$ term dominates as sequence length N grows, Compared to the baseline LLaVA [25] single forward pass where $N \approx 600$, SLoFo's main added cost comes from the second forward pass with teh initial sequence length $N \approx 1200$. With phase-wise pruning, SLoFo's second-pass overhead is estimated at 5.92 TMACs, compared to 8.15 TMACs ViCrop [40] which lacks token pruning. This confirms that the phase-wise token pruning strategy effectively reduces the overhead introduced by the additional sub-image.

Q: What is the color of the shovel? A. Yellow B. Red C. Blue D. Black

LLaVA-v1.5-7B: D ✗ SLoFo: C ✓

Q: What is the color of the Apple logo? A. polychromatic B. red C. white D. silver

LLaVA-v1.5-7B: D ✗ SLoFo: C ✓

Q: Is the umbrella on the left or right side of the traffic light? A. left B. right

LLaVA-v1.5-7B: A ✓ SLoFo: B ✗

Figure 7. Success(upper 2) and failure(lower 1) examples of SLoFo high-res on V*.