A. Appendix

A.1. Model Configurations

The LLaVA-7B model uses: (i) vision encoder, (ii) multi-layer perceptron (MLP) based image adapter/projector, and (iii) LLaMA 7B language model. The visual encoder is CLIP ViT-L/14 with details present in [16], the MLP-based image adapter has 2 linear layer with following sizes: 1024×4096 and 4096×4096. For the scenario when draft model also has image adapter the sizes are 1024×1024 and 1024×1024.

The following configurations are used for our target and draft language model part which follows the LLaMA architecture:

Table 1. Draft and target model configurations

	target (7B)	draft (115M)
Layers	32	4
Attention heads	32	8
Intermediate dim	11,008	2,816
Hidden dim	2,048	1,024
Activation	SiLU	SiLU

A.2. System Prompts

We use the following systems prompts for the respective task. The special image token is used to include the image data (*<image>*)

LLaVA-eval. We follow the prompt style given in [13], LLaVA has multiple questions and responses which we divide into different samples.

<s> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: <image>
Question Q_1 ASSISTANT: response R_1. USER: Question Q_2 ....

COCO-caption. As COCO dataset doesn't have any question prompts, we prompted the model with a prompt similar to above.

<s> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: <image>
Provide a detailed description of the given image ASSISTANT:

Science QA. We follow the prompt style provided in [14] with a single in-context example of the question, choices, answer and reasoning to enable Chain-of-Thought (CoT) reasoning. Additionally we only consider the test samples which have an associated image.

Question: question :I_i^{ques}
Options: (0) option : I_{i1}^{opt} (1) option : I_{i2}^{opt} (2) option : I_{i3}^{opt}
Context: context : I_i^{cont}
Answer: The answer is I_i^{ans}. BECAUSE: lectureI_i^{lect} explanation : I_i^{exp}

$< image >$
Question: question :I_{test}^{ques}
Options: (0) option : $I_{test,1}^{opt}$ (1) option : $I_{test,2}^{opt}$ (2) option : $I_{test,3}^{opt}$
Context: context : I_{test}^{cont}
Answer: The answer is

where, the subscript i is for in-context example.

In the SQA paper, the context field is provided by generating a caption for the associated image using an image captioning model, however, these captions were often simple and didn't provide a detailed description of the image

which is needed for answering the question. For this reason, the context field is filled with "hint" field provided in the SQA dataset. For the in-context sample we choose a sample without any associated image as the target LLaVA 7B cannot consume multiple images. We leave it as a future work to experiment SPD with more than 1 in-context examples.