Vision Inference Former: Sustaining Visual Consistency in Multimodal Large Language Models

Supplementary Material

7. Experimental details

In this section, we present the specific experimental setting including benchmarks and train details.

7.1. Benchmarks

We conducted extensive evaluations across a comprehensive set of 14 benchmark datasets, includin MMMU [45], RealWorldQA [43], MMBench [22], MMStar [4], OK-VQA [24], GQA [13], ScienceQA [32], and MMVP [39], OCRBench [21], TextVQA [34], AI2D [12], InfographicVQA [27], DocVQA [26] and POPE [19].

For DocVQA and InfographicVQA, we adopt the Average Normalized Levenshtein Similarity (ANLS) metric [46], while for all other datasets, we report standard accuracy metrics.

7.2. Train details

All experiments are conducted on a high-performance server equipped with eight NVIDIA H200 GPUs, each with 140 GB of memory. We employ DeepSpeed to enable efficient distributed training and memory optimization. The learning rate follows a cosine decay schedule with a 3% warm-up ratio, ensuring stable convergence during the early training phase. The maximum sequence length is set to 4096 tokens to accommodate long-context multimodal interactions.

Detailed hyperparameter configurations for training LLaVA and Qwen2.5-VL are provided in Table 6 and Table 7, respectively.

Our training pipeline consists of three sequential stages designed to ensure stable optimization and effective convergence: (1) Warm-up stage. We first train only the proposed Vision Inference Former and the LLM head on a small-scale pretraining dataset, while freezing all other parameters. This stage serves to initialize the newly introduced components and stabilize the optimization dynamics in the early phase. (2) Pretraining stage. Next, we conduct full-model fine-tuning on the complete multimodal corpus, enabling the model to achieve global adaptation across both visual and textual modalities. (3) Instruction tuning stage. Following the protocol of Ross [40], we further refine the model using a subset of the Cambrian dataset [37] to strengthen instruction-following capability while mitigating the risk of data leakage.

The pretraining corpus comprises LLaVA-Pretrain [20] and ShareGPT4V [3], which provide broad visual-language coverage for model initialization.

Table 6. Hyperparameters of training LLaVA.

Config	Step1	Step2	Step3
Trainable parts	VIF+LLM head	ALL	ALL
Global batch size	128	128	256
Batch size per GPU	8	8	16
Global learning rate	1e-4	2e-5	2e-5
Former learning rate	1e-4	4e-5	4e-5
Accumulated steps	2		
DeepSpeed zero stage	3		
Learning rate schedule	warmup + cosine decay		
Warmup ratio	0.03		
Weight decay	0		
Epoch	1		
Optimizer	AdamW		
Precision	bfloat16		
Model max length	4096		

Table 7. Hyperparameters of training Qwen2.5-VL.

Config	Step1	Step2	Step3
Trainable parts	VIF+LLM head	ALL	ALL
Global batch size	128	128	256
Batch size per GPU	8	8	16
Global learning rate	1e-4	1e-5	1e-5
Former learning rate	1e-4	2e-5	2e-5
Accumulated steps	2		
DeepSpeed zero stage	3		
Learning rate schedule	warmup + cosine decay		
Warmup ratio	0.03		
Weight decay	0		
Epoch	1		
Optimizer	AdamW		
Precision	bfloat16		
Model max length	4096		

The instruction tuning dataset includes LLaVA-Instruct, VQAv2 [11], GQA [13], OCRVQA [28], TextVQA [34], DVQA [15], DocVQA [26], ChartQA [25], ScienceQA [32], and MathVision [41].

For GQA, OCRVQA, TextVQA, DocVQA, and ScienceQA, we use the training splits for instruction tuning.

Table 8. Details of the instruction tuning dataset.

Method	Samples
LLaVA-Instruct	665k
VQAv2	240k
GQA_train	700k
OCRVQA_train	80k
TextVQA_train	34k
DVQA	39k
DocVQA_train	20k
ChartQA	2.5k
ScienccQA_train	6k
MathVision	3k

Table 9. Details of the pre-train dataset.

Method	Samples
LLaVA-Pretrain	558k
ShareGPT4V	80k

8. Case study

To qualitatively assess how the proposed Vision Inference Former (VIF) enhances visual grounding and reasoning consistency, we present representative case studies comparing LLaVA-1.5-7B and our LLaVA-VIF on visual question answering and image description tasks.

As shown in Figure 5, the question asks: "Where is the woman's blue bag located in the image?" The baseline LLaVA-1.5-7B predicts "In her hand," whereas LLaVA-VIF correctly answers "On her shoulder." This case exemplifies a common failure of connector-based models—the model's attention gradually shifts toward linguistic priors (e.g., the frequent co-occurrence of "hand" and "bag") instead of true visual evidence. By continuously injecting visual semantics into the decoding hidden states, VIF maintains stable alignment between the generated representation and the underlying visual features, resulting in an accurate and visually grounded answer.

In a free-form image description task (Figure 6), the baseline LLaVA-1.5-7B correctly described the image content in the early stages, but deviated in the latter half. In contrast, LLaVA-VIF identifies these details and provides a coherent, context-aware narrative of the scene. This improvement demonstrates VIF's ability to reinforce high-level semantic integration by directly linking decoding hidden states to uncompressed visual representations, thus preventing the loss of contextual cues during generation.

User:
Where is the woman's blue bag located in the image?
Options:
A: In her hand
B: On her shoulder
C: On the ground
D: Inside the man's bag

LLaVA: **A**
LLaVA-VIF: **B**

Figure 5. Case study in MMStar.

User: Describe this image.

LLaVA : The image features a sidewalk with a row of orange and white traffic cones placed along it. The cones are positioned in a straight line, creating a barrier to direct pedestrian traffic. There are a total of nine cones in the scene, with some closer to the foreground and others further back. In the background, there are two cars parked on the street, one on the left side and the other on the right side of the image. Additionally, there is a person standing near the middle of the scene, possibly observing the traffic cones or walking by.

LLaVA-VIF: The image captures a scene of a construction site or road work area. The main focus is a row of orange and white traffic cones, which are lined up along a sidewalk. These cones are connected by yellow caution tape, creating a barrier to guide pedestrians and vehicles safely around the work zone. The background of the image reveals a house and a tree, providing a sense of the location and context of the construction site. The overall scene suggests ongoing work or maintenance in the area, with the cones and tape serving as a visual indication of the ongoing activity.

Figure 6. Case study.