

Layer-wise Vision Injection with Disentangled Attention for Efficient LVLMs

Supplementary Material

In the supplementary materials, the following sections are included:

- **Implementation Details in Section A.**

This section provides a comprehensive description of the experimental setup, including detailed parameter settings. The aim is to ensure the reproducibility and clarity of the experiments.

- **Additional Experimental Results in Section B.**

This section consists of two parts: a detailed comparison of other methods and LVIDA, and an in-depth breakdown of LVIDA’s computational costs.

- **Qualitative Results in Section C.**

This section showcases detailed visualizations of the performance of LVIDA across various tasks.

A. Implementation Details

Table 1 summarizes the detailed hyperparameters used during training. All experiments are conducted on NVIDIA A100 GPUs.

Table 1. Hyperparameter Settings for Pretrain and Finetune

Parameter	Pretrain	Finetune
Training Modules	Vision Proj	Language Decoder
Learning Rate	1e-3	2e-5
Batch Size	256	128
LR Schedule	Cosine decay	
Optimizer	AdamW	
Weight Decay	0	
Zero Stage	Zero 3	
Warmup Ratio	0.03	
Data Precision	bf16	
Attention	Flash Attention 2	

B. Additional Results

B.1. Complementary Experiment on 8B Model

We conducted additional experiments using Llama3-8B [2] as the decoder, maintaining the same settings and training strategies as those in the main paper based on Vicuna-7B. Table 2 compares the results of the original model, LVIDA-7B and LVIDA-8B, demonstrating that LVIDA achieves superior performance while significantly reducing computational load across different normal-scale models.

B.2. Comparison with Other Methods

In the main paper, we introduced FastV [1] as a baseline to compare with LVIDA. FastV proposes a dynamic image token pruning method that reduces LVLM inference costs by removing certain visual tokens, whereas LVIDA achieves computational optimization by avoiding the forward propagation of the entire visual sequence within the language decoder.

In this subsection, we conduct a comprehensive comparison of FastV’s computational efficiency and performance under different parameter settings, using a model with Llama3-1B as the decoder. As shown in Table 3, although FastV achieves some reduction in computational cost, its overall performance declines significantly as the pruning intensity increases (i.e., removing more visual tokens). Even in its optimal configuration (filtering layer $K = 5$, visual token filtering rate $R = 50\%$), FastV’s FLOPs remain at 65% of the original model. In stark contrast, LVIDA reduces FLOPs to just 9% of the original model while maintaining performance comparable to the baseline, achieving a significant improvement in computational efficiency. LVIDA demonstrates clear advantages in both model performance and efficiency, underscoring its potential in real-world applications where both accuracy and computational cost are critical.

B.3. Detailed Computational Analysis

To provide a more in-depth analysis of the efficiency results presented in the main paper, this subsection examines the computational cost breakdown within the language decoder under different input configurations. Specifically, Table 4 presents the FLOPs (in GFLOPs) consumed by the two major components of the language decoder—attention and feed-forward networks (FFN). The V:L ratio represents the lengths of the vision and language sequences in the input prompt. Computational costs are reported for both the baseline models and the proposed LVIDA.

As described in Equation 7 of the main paper, LVIDA avoids the quadratic complexity of Vanilla-LVLM in the attention mechanism. For the FFN, the vision sequence is excluded entirely from computation, which significantly reduces the overall computational cost.

The results highlight LVIDA’s ability to optimize the computational cost of the language decoder while preserving performance metrics, positioning it as a practical solution for deploying LVLMs in resource-constrained scenarios with a strong balance between accuracy and efficiency. To further validate the robustness of LVIDA under vari-

Table 2. **Comprehensive Comparison of LVIDA and Baseline Models.** The V-L input ratio in these LVLMs is 728:64. The table reports GFLOPs and performance, highlighting the efficiency of LVIDA with comparable results.

Method	Language Decoder	FLOPs (G)	VQAv2	GQA	TextVQA	MM-Vet	POPE	MME	MMMU
LLaVA-SigLIP	Vicuna-7B	10800	80.6	63.3	63.7	40.0	86.6	1439.1	36.4
LVIDA	Vicuna-7B	1310	80.3	62.4	61.4	37.5	87.4	1498.9	38.3
LVIDA	Llama3-8B	1100	80.9	63.4	61.7	38.9	85.69	1498.6	39.2

Table 3. Performance and computational efficiency comparison of Llama3-1B, FastV (under different configurations with K for filtering layer and R for filtering ratio), and LVIDA. FLOPs are computed under an input configuration of V:L=728:64.

Language Decoder	VQA2	GQA	TextVQA	MM-Vet	POPE	MMMU	GFLOPs	Ratio
Llama3-1B	76.8	59.6	52.7	27.8	86.7	30.6	2040	100%
w.FastV (K=2 R=90%)	62.9	51.0	42.6	17.4	71.9	30.0	510	25%
w.FastV (K=2 R=75%)	72.2	56.3	48.4	24.5	84.1	29.8	758	37%
w.FastV (K=2 R=50%)	75.4	58.9	50.7	26.5	87.6	30.8	1180	58%
w.FastV (K=3 R=90%)	60.5	49.2	38.8	17.7	69.6	30.2	595	29%
w.FastV (K=3 R=75%)	70.8	55.1	45.5	24.0	82.8	30.3	829	41%
w.FastV (K=3 R=50%)	75.0	58.5	49.7	27.7	87.0	30.8	1230	60%
w.FastV (K=5 R=90%)	65.8	51.5	39.9	18.3	73.6	30.0	765	38%
w.FastV (K=5 R=50%)	75.6	59.0	50.3	28.6	87.2	30.8	1320	65%
w.LVIDA	76.2	59.6	50.5	28.4	86.6	29.9	192	9%

Table 4. Component-wise Computational Costs (FLOPs) of the Language Decoder Across Different V:L Ratios. This table provides a detailed breakdown of the FLOPs consumed by attention (denoted as F_Attn) and feed-forward networks (denoted as F_FFN) within the language decoder for different vision-to-language input length ratios.

Method	V:L 728:32		V:L 728:64		V:L 728:200		V:L 728:728		V:L 728:1000	
	F_Attn	F_FFN	F_Attn	F_FFN	F_Attn	F_FFN	F_Attn	F_FFN	F_Attn	F_FFN
Qwen2-0.5B	117	476	124	497	156	582	311	914	410	1085
LVIDA	15	19	20	40	44	126	166	457	248	628
TinyLlama-1.1B	420	1157	442	1206	541	1413	988	2217	1258	2631
LVIDA	37	49	55	97	136	304	513	1109	747	1523
Llama-3.2-1B	331	1224	348	1276	425	1495	768	2345	973	2783
LVIDA	41	50	56	103	119	322	411	1173	590	1611
Llama-3.2-3B	1270	3213	1333	3349	1605	3924	2783	6156	3465	7306
LVIDA	148	131	204	270	442	846	1488	3078	2101	4228

ous input configurations, we conduct two additional experiments: (1) model efficiency under longer language sequences, and (2) total inference latency including the vision encoder.

Table 5 reports FLOPs, latency (TTFT), and peak memory consumption under extended V:L ratios (728:728 and 728:1000). The results show that LVIDA consistently achieves significant reductions in all metrics, even under high input loads.

Table 6 presents both decoder-side and total end-to-end inference latency across different input lengths. LVIDA maintains clear efficiency gains in total inference time, highlighting its practical benefits beyond the language decoder alone.

Table 5. Model efficiency under different V:L ratios.

Language Decoder	V:L 728:728			V:L 728:1000		
	FLOPs	TTFT	Memory	FLOPs	TTFT	Memory
Llama-3B	10090	228.7	24.3	12130	287.9	28.9
LVIDA	5140	125.1	17.0	7120	177.5	20.9

Table 6. Decoder and total TTFT under varying input ratios.

Language Decoder	V:L 728:64		V:L 728:200	
	Decoder	Total	Decoder	Total
Llama-3.2-3B	119.8	142.3	143.1	164.4
LVIDA	37.6	65.3	49.9	76.8
Vicuna-7B	215.6	240.1	250.5	274.4
LVIDA	54.4	89.1	82.7	127.2

The tool used for testing FLOPs is [3].

C. Qualitative Results

This section provides qualitative results to offer a more intuitive understanding and comparison between the baseline model (e.g., Llama3-1B) and LVIDA. Specifically, we showcase multiple tasks, including choice questions (Figure 1), yes/no questions (Figure 2), simple image captions and detailed image captions (Figures 3, 4, and 5), object recognition (Figure 6). We observe the following results:

- LVIDA achieves performance comparable to the baseline. Across various multimodal tasks, LVIDA accurately understands and responds to the requirements, demonstrating strong alignment with task-specific objectives.
- When the task requires detailed image descriptions, both the baseline model and LVIDA occasionally generate similar hallucinated descriptions (highlighted in red in the images). We attribute this behavior to the inherent limitations of the selected baseline model. Nevertheless, these results confirm that LVIDA does not compromise the baseline model’s ability to process and interpret vision information, maintaining performance parity with the baseline model.

References

- [1] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. 1
- [2] 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 preprint arXiv:2407.21783*, 2024. 1
- [3] xiaoju ye. calflops: a flops and params calculate tool for neural networks in pytorch framework, 2023. 3


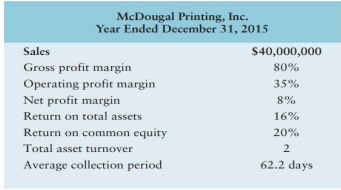

Choice Questions		
 <p>Question: What is the title of this design?</p> <p>(A) Zenith Guardian Ear & Radio Nurse (B) Innovative Hearing Aid Technology (C) Ear and Radio Communication Advancements (D) The Future of Guardian Technology</p> <p>Llama3-1B: A</p> <p>LVIDA: A</p>	 <p>Question: McDougal, Inc., had sales totaling \$40,000,000 in fiscal year 2015. Some ratios for the company are listed below. Use this information to calculate values for the operating profits.</p> <p>(A) \$14,000,000 (B) \$8,000,000 (C) \$32,000,000 (D) \$18,000,000</p> <p>Llama3-1B: A</p> <p>LVIDA: A</p>	 <p>Question: Identify the biome shown in image</p> <p>(A) taiga (B) tundra (C) rain forest (D) desert</p> <p>Llama3-1B: B</p> <p>LVIDA: B</p>

Figure 1. Qualitative results of choice questions.


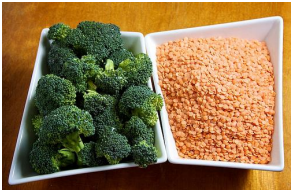

Yes-or-No Questions		
 <p>Question: Is there a person in the image?</p> <p>Llama3-1B: Yes</p> <p>LVIDA: Yes</p> <p>Question: Is there a dining table in the image?</p> <p>Llama3-1B: No</p> <p>LVIDA: No</p>	 <p>Question: Is there a broccoli in the image?</p> <p>Llama3-1B: Yes</p> <p>LVIDA: Yes</p> <p>Question: Is there a dining table in the image?</p> <p>Llama3-1B: Yes</p> <p>LVIDA: No</p>	 <p>Question: Is there a dog in the image?</p> <p>Llama3-1B: No</p> <p>LVIDA: No</p> <p>Question: Is there a tennis racket in the image?</p> <p>Llama3-1B: No</p> <p>LVIDA: No</p>

Figure 2. Qualitative results of yes/no questions. Correct answers and Wrong answers are highlighted in color respectively.

Image Caption



Generate a simple caption for this image.

Llama3-1B: A group of **children playing** a video game in a living room.

LVIDA: A group of four **people are playing** a video game in a living room.

Generate a detailed description of this image.

Llama3-1B: The image depicts a group of four young boys **playing a video game** together in a living room. They are standing in front of a window, which is partially open, **allowing natural light to enter the room**. The boys are **holding game controllers**, actively engaged in the game. In the room, there are two chairs. A couch can be seen in the background, occupying the right side of the room.

LVIDA: The image shows a group of four people, including two boys and two girls, standing in a living room and **playing a video game** together. They are all holding **Wii controllers** and appear to be enjoying the game. The living room is furnished with a couch and a couple of chairs, providing a comfortable space for the group to play and socialize.



Generate a simple caption for this image.

Llama3-1B: A man **holding a fish** in front of a river.

LVIDA: A man wearing a hat and a suit is **holding a fish** he caught.

Generate a detailed description of this image.

Llama3-1B: The image features a man **wearing a hat and a suit**, standing by a river with a large **fish in his hand**. He is holding the fish **with pride**, showcasing his catch. The man appears to be a fisherman, possibly posing for a picture with the fish he has just caught.

LVIDA: The image features a man **wearing a hat and a suit**, **holding a large fish** he has caught in his hands. He is posing for a picture while **proudly** displaying his catch. The man is standing near a body of water, which could be a river or a lake, as he is holding a fish that he caught in that location.

Figure 3. Qualitative comparison of image captions. **Correct descriptions** and **hallucinated** are highlighted in color respectively.

Image Caption



Generate a simple caption for this image.

Llama3-1B: A soccer player in a red and white uniform is jumping in the air.

LVIDA: A soccer player wearing red and white jumps in mid air.

Generate a detailed description of this image.

Llama3-1B: The image captures a lively soccer game in progress, with a player in red and white uniform leaping into the air to kick the ball. The player is in the middle of the field, surrounded by several other players and spectators. A sports ball can be seen in the middle of the field.

LVIDA: The image features a soccer field with a man wearing a red and white striped uniform, standing in the air and jumping to catch a soccer ball. He is in the process of making a powerful leap, showcasing his athleticism. In the background, there are several other people scattered around the field, possibly watching the game or waiting for their turn to play.



Generate a simple caption for this image.

Llama3-1B: A cat is sitting in a toilet bowl and looking at the camera.

LVIDA: A cat is sitting in a toilet bowl and looking up.

Generate a detailed description of this image.

Llama3-1B: The image features a brown and white cat sitting inside a toilet bowl. The cat appears to be enjoying its time in the toilet, as it is comfortably curled up and resting. The toilet bowl is filled with water, creating a unique and amusing scene for the cat.

LVIDA: The image features a large, fluffy orange and white cat sitting inside a toilet bowl. The cat appears to be relaxed and comfortable, with its paws up and resting on the toilet seat. The toilet bowl is filled with water, creating a unique scene of the cat enjoying its time in the bathroom.

Figure 4. Qualitative comparison of image captions. Correct descriptions and hallucinated are highlighted in color respectively.

Image Caption



Generate a simple caption for this image.

Llama3-1B: A dog is holding a hot dog in its mouth.

LVIDA: A dog is holding a hot dog in its mouth and wearing a collar.



Generate a simple caption for this image.

Llama3-1B: A man riding a surfboard in the ocean.

LVIDA: A surfer wearing a red shirt riding a wave on a surfboard.

Generate a detailed description of this image.

Llama3-1B: The image features a black and white dog sitting on a green chair, holding a hot dog in its mouth. The dog appears to be enjoying the hot dog, as it is smiling and chewing on it. The chair is positioned in the center of the scene, with the dog occupying the majority of the frame.

LVIDA: The image features a black and white dog wearing a red collar, lying on a green bean bag chair and enjoying a hot dog. The dog appears to be chewing on the hot dog, which is placed in its mouth. The scene is set in a cozy environment, with the dog comfortably resting on the bean bag chair.

Generate a detailed description of this image.

Llama3-1B: The image captures a thrilling moment of a man riding a wave on a surfboard. He is skillfully maneuvering through the water, showcasing his expertise in surfing. In addition to the main surfer, there are two other people visible in the scene. One person is located towards the left side of the image, and the other person is positioned slightly to the right of the main surfer. Both of them appear to be enjoying the ocean waves as well.

LVIDA: The image features a man riding a surfboard on a wave in the ocean. He is skillfully maneuvering his surfboard, making a turn as he rides the wave. The surfer is wearing a red shirt, which stands out against the blue ocean water. There are several other smaller waves in the scene, some of which are also being ridden by surfers. The surfers are spread out across the ocean, with some closer to the shore and others further out in the water.

Figure 5. Qualitative comparison of image captions. Correct descriptions and hallucinated are highlighted in color respectively.

Object Recognition		
		
<p>Question: What type of animal is shown in the image?</p> <p>Llama3-1B: The image shows a dog.</p> <p>LVIDA: The image shows a dog.</p>	<p>Question: What kind of animal is visible in the image?</p> <p>Llama3-1B: A donkey is visible in the image.</p> <p>LVIDA: A donkey is visible in the image.</p>	<p>Question: Can you describe the clothes worn by the people in the picture?</p> <p>Llama3-1B: One woman is wearing a blue shirt, another is wearing a white shirt, and the third woman is wearing a blue sweater.</p> <p>LVIDA: One the woman is wearing a white shirt, and the two other people are dressed in blue clothing.</p>
<p>Question: What are they playing?</p> <p>Llama3-1B: The woman and a brown dog are playing Frisbee.</p> <p>LVIDA: The woman and dog are playing Frisbee together in a grassy field.</p>	<p>Question: Can you identify the type of object being offered to the animal?</p> <p>Llama3-1B: The object being offered to the animal is a carrot.</p> <p>LVIDA: Yes, the object being offered to the animal is a carrot.</p>	<p>Question: Which person in the picture is carrying a bag? What colour is it?</p> <p>Llama3-1B: In the picture, the woman in the blue shirt is carrying a black bag.</p> <p>LVIDA: The person in the picture carrying a black bag is a woman wearing a white jacket.</p>

Figure 6. Qualitative results of object recognition tasks. **Correctly identified objects** and **errors** are highlighted in color respectively.