Visual Layer Selection Matters for Egocentric VLM Perception

Ruchen Liu[1,*], Yi Yang[1,*], Yiming Xu[1,*], Monika Sester[1], Bodo Rosenhahn[1]

***Abstract*— Vision-language models are increasingly used for first-person perception and reasoning, and are becoming relevant to robot perception in unstructured environments. Existing methods typically use a fixed visual layer from the vision tower, but the reliability of this default choice is rarely examined. This work studies visual layer selection in egocentric VLM perception with a controlled layer-wise scanning protocol. For each candidate layer, only the source of visual features is replaced, while the rest of the multimodal inference pipeline remains unchanged. Experiments across multiple VLMs on egocentric video question answering tasks show that mid-to-high visual layers often outperform the default layer. The results indicate that the deepest representation is not always the most effective input for downstream reasoning. Patch-level activation analysis further shows that better-performing layers produce more spatially concentrated responses than the default and final layers. These findings identify visual representation depth as an important factor in egocentric VLM perception. They also suggest that layer-wise analysis provides useful interpretability signals for perception modules in end-to-end multimodal and robot-relevant systems.**

I. Introduction

Robots operating in unstructured environments require perception systems that can support reliable and interpretable decision making. Recent Vision-Language Models (VLMs) provide a promising foundation for such settings because they combine visual perception with language-based reasoning [1], [2]. This is especially relevant for first-person perception tasks, where the system must understand object interactions, spatial relations, and action-related cues from complex visual scenes. Many current VLM-based pipelines use a fixed visual layer from the vision backbone, i.e., the vision tower, as the input to downstream reasoning. However, the reliability of this default choice is rarely examined, despite the hierarchical nature of vision transformers and prior layer-wise analyses in visual representation learning [3], [4]. As a result, an important design factor in end-to-end multimodal perception remains insufficiently understood.

Prior studies have shown that vision transformers form different types of visual representations across depth. Early layers emphasize local patterns and low-level visual content, intermediate layers progressively integrate spatial structure and semantic cues, and deeper layers become more abstract [3], [5]. However, these observations have mainly been established in single-modality vision settings, such as image classification and representation probing [4]. In multimodal systems, the vision encoder does not produce the final prediction directly. Instead, its output is provided to the language model as the visual representation for downstream reasoning. As a result, layer-wise patterns observed in visual-only tasks do not directly determine which representation depth is most effective for multimodal reasoning.

To study this question, this work adopts a controlled layer-wise scanning protocol to evaluate different depths of the vision tower. Only the source of visual features is replaced, while the rest of the multimodal inference pipeline remains unchanged. Experiments on egocentric video question answering across multiple VLMs show that mid-to-high visual layers often outperform the default layer. These results indicate that visual representation depth is an important factor in egocentric VLM perception.

[1]Leibniz University Hannover, Hannover, Germany.
*These authors contribute equally to this work.
Ruchen Liu: ruchen.liu@stud.uni-hannover.de
Yi Yang: yangyi@tnt.uni-hannover.de
Yiming Xu: yiming.xu@ikg.uni-hannover.de
Monika Sester: monika.sester@ikg.uni-hannover.de
Bodo Rosenhahn: rosenhahn@tnt.uni-hannover.de

II. Method

A video VLM typically consists of a vision tower, a projector, and a language model. Given an input video, the vision tower produces hidden representations at different depths. In standard inference, one predefined visual layer is selected as the feature source for downstream reasoning. As shown in Fig. 1, this work studies visual layer selection with a controlled layer-wise scanning protocol.

Let $\mathbf{H}^{(l)}$ denote the hidden representation from layer l of the vision tower. For each candidate layer, $\mathbf{H}^{(l)}$ is used as the visual input to the projector, replacing the default layer output. Only the source of visual features is changed. The rest of the multimodal pipeline remains unchanged, including the projector, the language model, the prompt format, the decoding strategy, and all pretrained parameters.

For every candidate layer, full multimodal inference is preserved. The entire vision tower is executed, and no early-exit or architectural modification is introduced. This design ensures comparison of different representation depths under the same end-to-end reasoning setting. Therefore, differences across layers can be attributed to representation depth rather than changes in model structure or computation.

III. Experiments

Experiments are conducted on HD-EPIC VQA tasks [6] using multiple VLMs. HD-EPIC provides first-person video question answering tasks with fine-grained visual and interaction cues, making it a suitable benchmark for studying layer selection in egocentric multimodal perception. The task pool considered in this work includes Gaze Estimation (1000 samples), How Recognition (500), Why Recognition (500),

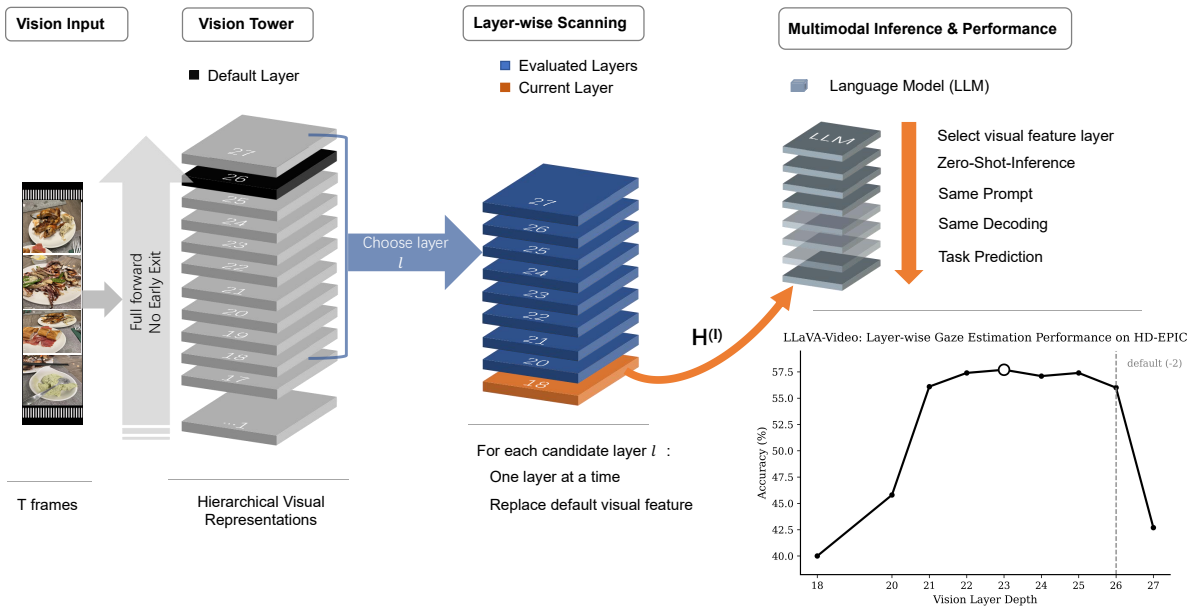

Fig. 1: Overview of the controlled layer-wise scanning protocol. Different vision tower layers are evaluated by replacing only the visual feature source $\mathbf{H}^{(l)}$, while the rest of the multimodal inference pipeline remains unchanged.

Ingredient Weight (50), and Action Recognition (200 uniformly sampled examples). The main evaluation is performed on the state-of-the-art open-source VLM LLaVA-Video [7], with additional validation on LLaVA-NeXT [8] and Video-LLaVA [9]. LLaVA-Video uses a 27-layer SigLIP vision tower [10], while LLaVA-NeXT and Video-LLaVA use 24-layer CLIP-based vision towers [11].

For each model, candidate layers from the middle-to-upper part of the vision tower are evaluated by replacing the default visual feature layer, while the rest of the inference pipeline remains unchanged. To ensure fair comparison across layers, all scanned layers within the same model are evaluated with the same prompt format, decoding strategy, frame sampling configuration, and evaluation protocol. No parameter update or task-specific finetuning is introduced, so performance differences across layers can be attributed to representation depth under the same end-to-end inference setting.

LLaVA-Video and LLaVA-NeXT are evaluated with 16 uniformly sampled frames, while Video-LLaVA is evaluated with at most 8 uniformly sampled frames due to its model configuration. This difference in frame budget also provides a supplementary check of whether the observed layer-wise trend is sensitive to temporal input granularity. Despite the lower frame limit, the layer-scan behavior remains comparable, which supports the stability of the main observation.

IV. Results and Discussion

A. Layer-wise Performance Across Models

Across models and egocentric tasks, a consistent layer-wise trend is observed. As shown in Figs. 2–4, performance often improves from lower candidate layers to the middle-to-upper part of the vision tower, reaches its best region before or around the default output layer, and then drops at the deepest layer. A similar pattern is observed across different VLMs despite differences in backbone architecture, default layer choice, task coverage, and temporal input budget. The results indicate that the default visual layer is not consistently the most effective choice for downstream reasoning. Instead, intermediate upper layers often provide stronger visual inputs for egocentric VLM perception.

Figure 2 shows clearly and stably such a trend on LLaVA-Video. On Gaze Estimation, performance rises sharply from lower candidate layers and reaches a high plateau around layers 21–26, with the best result appearing before the default layer. Similar behavior is also observed on Why Recognition and Ingredient Weight, where performance continues to improve toward the middle-to-upper layers and peaks at layer 25 rather than the default layer 26. How Recognition is slightly different in that the default layer remains competitive, but the best result still appears one layer earlier. Overall, the LLaVA-Video curves suggest that the middle-to-upper region of the SigLIP vision tower provides a more favorable representation range than the default output layer.

A similar tendency is also observed on LLaVA-NeXT in Fig. 3. Although the task set is not identical to that of LLaVA-Video, the curves again show that the best performance often appears before the default layer. This is especially clear for How Recognition and Action Recognition, where the optimal layers are located several layers earlier than the default output. Gaze Estimation is more gradual

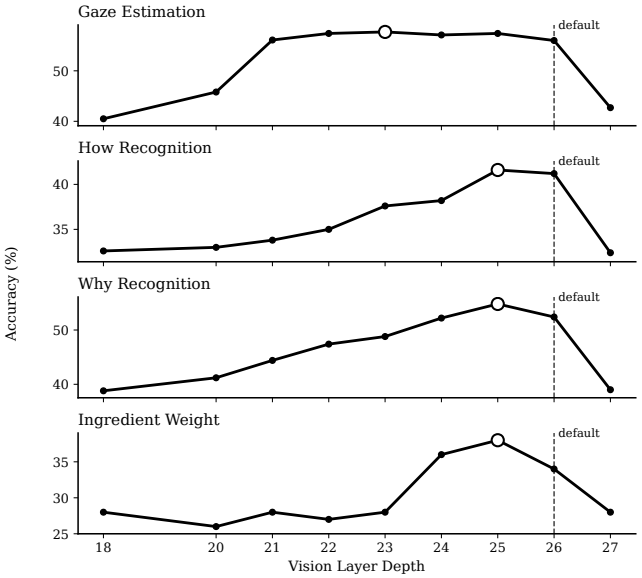

Fig. 2: Layer-wise performance of LLaVA-Video on HD-EPIC. Performance generally improves toward the middle-to-upper layers and peaks before the default output layer.

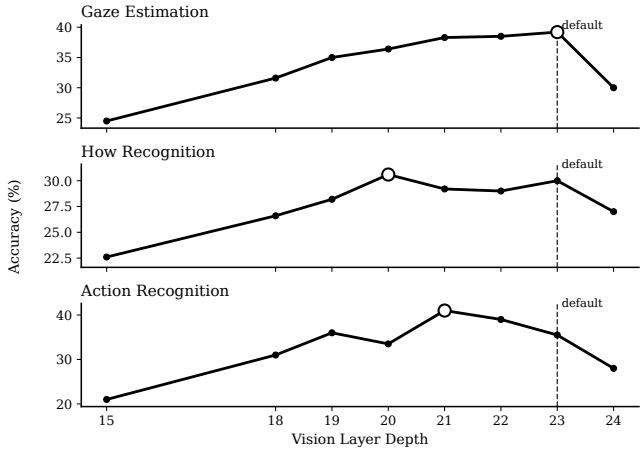

Fig. 3: Layer-wise performance of LLaVA-NeXT on HD-EPIC. A similar trend is observed, with intermediate upper layers often outperforming the default layer.

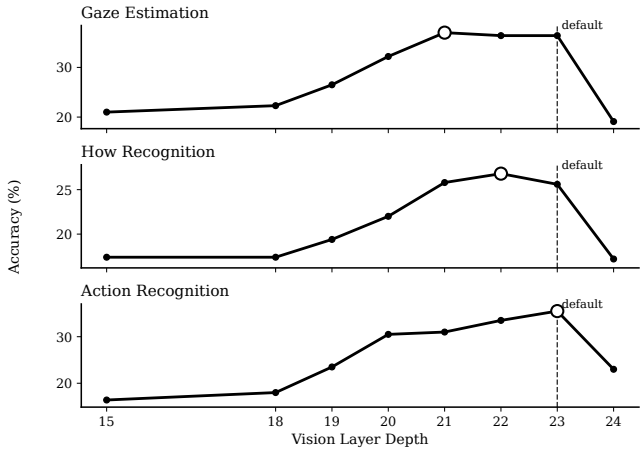

Fig. 4: Layer-wise performance of Video-LLaVA on HD-EPIC. Despite the lower frame budget, the layer-wise trend remains comparable.

and the default layer remains strong, but performance still does not improve further at the deepest layer. These results show that the layer-wise effect is not limited to one model or one task configuration.

Figure 4 further supports this observation under a more constrained temporal setting. Due to the model configuration, Video-LLaVA is evaluated with at most 8 input frames. Even under this lower frame budget, the same general pattern remains visible. Gaze Estimation reaches its best result before the default layer, and How Recognition also peaks earlier than the default output. Action Recognition is the only task where the default layer remains the strongest, but the curve still shows a clear rise from lower layers to the middle-to-upper region, followed by a drop at the final layer. This suggests that the main layer-wise trend is not tied to a single frame setting.

Taken together, these results support two observations. First, the default visual layer is often competitive, but it is not consistently optimal across models and tasks. Second, the most effective visual input for egocentric multimodal reasoning often comes from the middle-to-upper part of the vision tower rather than from its deepest available representation.

Although HD-EPIC is not a robotics benchmark, its egocentric VQA tasks probe several perception abilities that are closely relevant to robot systems. Gaze Estimation is related to task-relevant object grounding and selective attention under cluttered views. How Recognition and Action Recognition are related to interaction understanding and embodied action perception. Why Recognition further introduces goal- and context-aware reasoning. Together, these tasks provide a useful first-person testbed for studying how visual representation depth affects robust and interpretable multimodal perception.

B. Patch-Level Spatial Response Analysis

To support our finding that intermediate upper layers often perform better, we further find empirical clues from patch-level activation patterns, examined on LLaVA-Video for the Gaze Estimation task. A clear trend emerges from the spatial distribution of visual responses. Compared with the default and final layers, better-performing intermediate layers tend to produce more spatially concentrated activations and less diffuse global responses. In other words, the visual evidence used for downstream reasoning becomes more localized in space.

A closer inspection is first conducted on two samples randomly selected from the 1000 Gaze Estimation examples. As shown in Table I, layer 23 (optimal) consistently exhibits lower mean energy than both layer 26 (default) and layer 27 (final). At the same time, its entropy remains comparable to or lower than that of the default and final layers. These sample-level results suggest that the stronger layer does

TABLE I: Patch-level energy statistics for two samples randomly selected from the 1000 Gaze Estimation examples. Layer 23 is the optimal layer, layer 26 is the default layer, and layer 27 is the final layer.

Sample	Layer	Mean	Concentration	Entropy
S171	27 (final)	66.93	0.1467	0.9955
S171	26 (default)	72.83	0.1337	0.9977
S171	23 (optimal)	46.68	0.1438	0.9957
S204	27 (final)	66.32	0.1497	0.9948
S204	26 (default)	76.75	0.1317	0.9981
S204	23 (optimal)	51.49	0.1375	0.9971

TABLE II: Full-dataset patch-level energy statistics on the 1000 Gaze Estimation examples. The comparison is made between layer 23 (optimal) and layer 26 (default). Δ is computed as $(\mathrm{opt} - \mathrm{def})$.

Metric	Avg. Δ (opt$-$def)	Consistency
Mean energy	$-25.32 \downarrow$	1000/1000
Normalized entropy	$-0.0019 \downarrow$	1000/1000
Gini coefficient	$+0.0234 \uparrow$	993/1000

not simply amplify all responses, but redistributes activation more selectively across spatial locations.

The same pattern is further supported at the dataset level. As shown in Table II, the optimal layer shows lower mean energy and lower normalized entropy than the default layer across all 1000 evaluated samples, while also yielding a higher Gini coefficient in 993 out of 1000 cases. Here, the Gini coefficient measures how unevenly the patch-level energy is distributed across spatial locations. A higher value indicates that activation energy is concentrated on fewer patches rather than spread more uniformly across the image. Since entropy and Gini are computed on normalized spatial energy distributions, these changes indicate a systematic redistribution of spatial response rather than a trivial global scaling.

These observations do not by themselves establish a full causal explanation, but they provide an interpretable signal for why representation depth affects performance. The advantage of intermediate upper layers may come from a more favorable balance between semantic abstraction and spatial specificity. This supports the view that layer-wise analysis can serve not only as a performance study, but also as a useful tool for understanding perception quality in end-to-end multimodal systems.

V. Conclusion

This work studies visual layer selection in egocentric VLM perception with a controlled layer-wise scanning protocol. Across multiple VLMs and HD-EPIC VQA tasks, intermediate upper layers often outperform the default visual layer under the same end-to-end inference setting. Additional analysis of patch-level activation patterns shows that stronger layers tend to produce more spatially concentrated responses than the default and final layers. These findings identify representation depth as an important factor in egocentric multimodal perception, and suggest that layer-wise analysis can provide a useful perspective for understanding perception quality in robot-relevant first-person reasoning settings.

VI. Limitations and Future Work

This study focuses on egocentric VQA rather than closed-loop robotic control, evaluates models in a zero-shot setting without task-specific finetuning, and restricts the scan to a predefined middle-to-upper subset of the vision tower. Although consistent trends emerge across models and tasks, broader validation on embodied benchmarks remains necessary for future research and applications.

Nevertheless, the observed layer-wise behavior carries implications for robot perception pipelines, where visual representations support action planning, object interaction, and task-relevant attention under dynamic first-person observations. Our finding that intermediate upper layers yield more effective and spatially concentrated representations suggests that visual layer selection may affect not only multimodal reasoning but also the reliability and interpretability of embodied perception modules. We thus view this work as an initial step toward depth-aware perception design for robot-relevant multimodal architectures. Future work may explore adaptive layer selection and depth-aware routing, and examine whether these trends persist under supervised adaptation, larger-scale MLLMs, and embodied manipulation and navigation benchmarks.

References

[1] H. Liu *et al.*, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[3] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[4] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2018. [Online]. Available: https://arxiv.org/abs/1610.01644

[5] M. Raghu *et al.*, "Do vision transformers see like convolutional neural networks?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] T. Perrett *et al.*, "Hd-epic: High-definition egocentric video understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[7] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," 2024. [Online]. Available: https://arxiv.org/abs/2410.02713

[8] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, "Llava-next: A strong zero-shot video understanding model," April 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[9] B. Lin *et al.*, "Video-llava: Learning unified video-language understanding," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[10] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[11] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.