Supplementary Material for

LLMind: Bio-inspired Training-free Adaptive Visual Representations for Vision-Language Models

Soumyaratna Debnath Bui Duc Manh Zinan Liu Lin Wang
Nanyang Technological University, Singapore
soumyara004@e.ntu.edu.sg ducmanh.bui@ntu.edu.sg zinan001@e.ntu.edu.sg linwang@ntu.edu.sg

1. Additional Experimental Results

The pseudocode for the proposed Bio-Inspired Adaptive Sampling Strategy (BASS) is provided in Algorithm 1.

1.1. Details of the VLMs

Table 1 reports the parameter sizes of all VLMs evaluated in our experiments. For clarity, we list the model family, the exact variant used, and its total number of parameters.

Table 1. Parameter size of VLMs used in our experiments.

Model Family	Variant	Params (B)
Qwen2.5-VL [1]	Qwen2.5-VL-3B-Instruct	4.0
SmolVLM [7]	SmolVLM-Instruct	2.0
LLaVA-OneVision [5]	Llava-Onevision-Qwen2-0.5b	0.9
Qwen3-VL [9]	Qwen3-VL-2B-Instruct	2.0
DeepSeek-VL [6]	Deepseek-vl-1.3b-chat	2.0

1.2. Comparison on Scene-level VQA

Fig. 7 and Fig. 8 present additional qualitative results on the SEED-Bench dataset [4], illustrating the behavior of Qwen2.5-VL under 5% and 3% pixel budgets, respectively. Additional category-wise analysis shows that LLMind consistently improves performance across all major reasoning types, indicating broad, task-agnostic gains (Fig. 3).

Fig. 4 presents the loss curves for different VLMs on both the VQAv2 [3] and SEED-Bench [4] datasets across 1%, 3%, and 5% pixel budgets. Across all settings, Qwen2.5-VL, SmolVLM, and LLaVA-OneVision exhibit similar convergence behavior, with higher pixel budgets yielding lower and more stable loss trajectories.

Additionally, Fig. 5 presents the qualitative results for LLMind on the SEED-Bench dataset [4] for a **black-box** model, *Gemini-2.5-Flash* [2] under 5% pixel budget. The result demonstrates that the proposed LLMind method adapts seamlessly to both *white-box* and *black-box* VLMs without requiring model-level access.

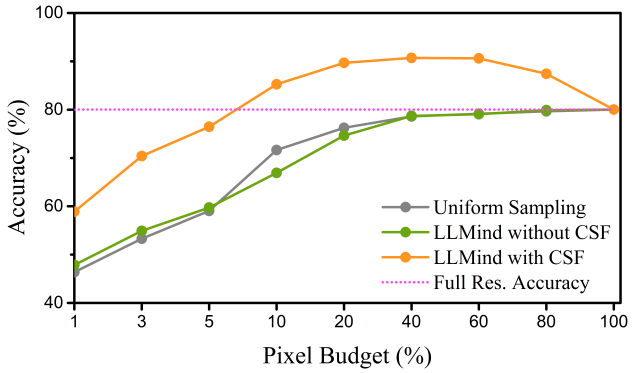


Figure 1. Ablation study on VQAv2 dataset with SmolVLM illustrating the impact of pixel density and the CSF module.

To better illustrate the optimization behavior of LLMind, we include a supplementary video (*LLMind.mp4*) that visualizes the progression of the optimization process.

1.3. Comparison on Region-guided VQA

Fig. 6 compares the distribution of correct, partially correct[1], and wrong predictions for Uniform Sampling and LLMind across different pixel budgets for multiple region-specific classification (RSC) based VQA tasks (*Refer to Table 3 of main draft*).

Multiple-RSC refers to the classification of multiple objects within the same image, where the model must correctly identify several region-grounded targets simultaneously, making the task more sensitive to information loss under low pixel budgets.

Across all the settings, LLMind produces a notably higher proportion of correct answers while consistently reducing both wrong and partially correct responses. The gains become more prominent as the pixel budget increases, with LLMind achieving up to **44%** correct responses at 10% budget, compared to 25% under Uniform Sampling.

[1]Partially correct means that the model answers some, but not all, of the questions correctly.

Table 2. **Ablation study on pixel density and the CSF module** for Qwen2.5-VL (A-OKVQA) and SmolVLM (VQAv2). The best accuracy in each column is shown in **bold**, while the second-best result is underlined. **LLMind with CSF** delivers the strongest performance across all budgets, highlighting the complementary benefits of dense sampling and semantic feedback.

Model	Dataset	Accuracy (%) Full Res.	Sampling Methods	Accuracy Sampled (%)								
				1%	3%	5%	10%	20%	40%	60%	80%	100%
Qwen2.5-VL	A-OKVQA	85.67	Uniform Sampling	44.31	48.77	52.01	68.23	76.83	81.21	83.11	83.51	85.67
			LLMind without CSF	44.46	49.79	52.65	64.74	71.06	78.05	84.46	84.66	≈
			LLMind with CSF	**51.18**	**61.53**	**69.74**	**83.28**	**89.35**	**90.98**	**90.49**	**89.16**	≈
SmolVLM	VQAv2	80.01	Uniform Sampling	46.38	53.28	59.06	71.64	76.24	78.61	79.15	79.66	80.01
			LLMind without CSF	47.85	54.93	59.75	66.92	74.64	78.69	79.07	79.90	≈
			LLMind with CSF	**58.88**	**70.40**	**76.46**	**85.24**	**89.67**	**90.69**	**90.60**	**87.41**	≈

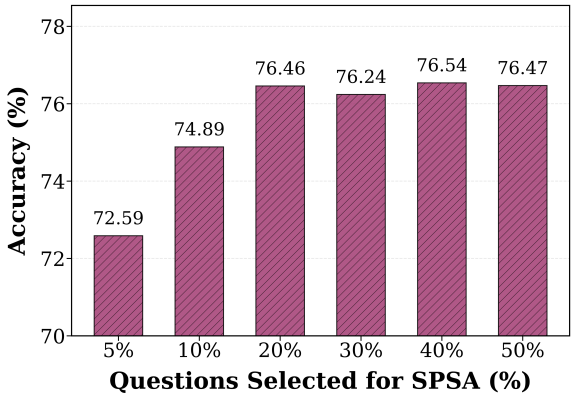


Figure 2. **Ablation of SPSA hyperparameters.** We study the influence of the question subsampling ratio (p) and related settings on performance.

1.4. Ablation and Analysis

Component Analysis. We conduct an ablation study on the A-OKVQA dataset using Qwen2.5-VL (*Refer to Fig. 7 of the main draft*), and on VQAv2 dataset using SmolVLM to analyze the roles of pixel density and the CSF module. Table 2 reports the quantitative effect of each component and Fig. 1 visualizes the performance trends across varying pixel budgets for VQAv2 dataset using SmolVLM.

Effect of Iterations. We further analyze the impact of the number of optimization iterations under a fixed 5% pixel budget using SmolVLM on VQAv2. We observe a consistent performance improvement as the number of iterations increases, with diminishing gains beyond 100 iterations: 64.62%@10 iters; 72.59%@50 iters; 76.46%@100 iters; 76.96%@150 iters; 76.99%@200 iters. This indicates that the CSF-driven refinement converges effectively within a relatively small number of iterations.

SPSA Question Subsampling. SPSA is performed using a random subset of p% questions to reduce computational overhead during optimization. We analyze the impact of the subset ratio p in Fig. 2, showing that a small subset is sufficient to achieve stable and competitive performance.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 4

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[4] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 4, 5, 6

[5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[6] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1

[7] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 1

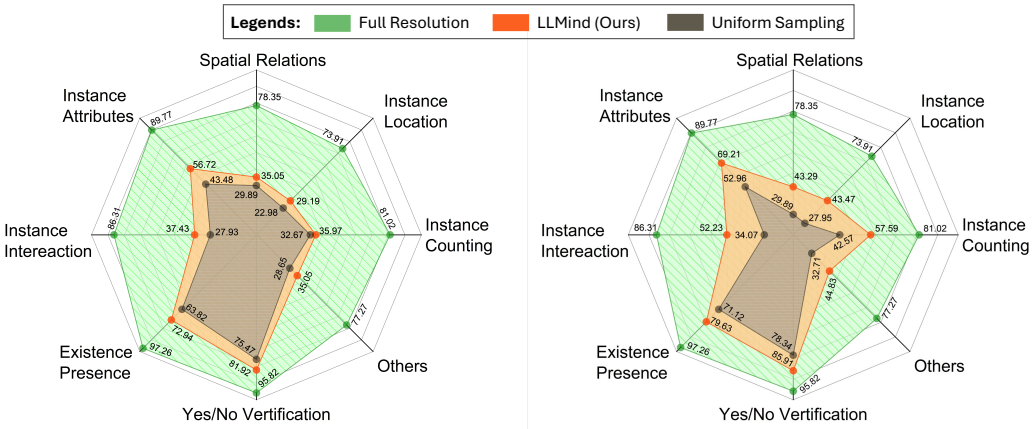
[8] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 3

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

Algorithm 1 BASS: Bio-Inspired Adaptive Sampling Strategy (*Refer to Fig. 4 of the main draft*)

Require: *params*: optimized transformation parameters from the MLP ; *percentage*: pixel budget

function BASS(params, percentage)
  /* 1. Forward Möbius Transform */
  $(a, b, c, d) \leftarrow$ GETPARAMS$(params)$
  warped_img $\leftarrow$ MOBIUSTRANSFORM(IMAGE, a, b, c, d) /* Eq. 3 (main draft) */
  /* 2. Uniform Sampling */
  sampled_img $\leftarrow$ UNIFORMSAMPLEANDINTERPOLATE(warped_img, percentage)
  /* 3. Inverse Möbius Transform */
  $(a^{-1}, b^{-1}, c^{-1}, d^{-1}) \leftarrow$ GETINVERSEPARAMS$(params)$
  reconstructed $\leftarrow$ INVERSEMOBIUSTRANSFORM(sampled_img, $a^{-1}, b^{-1}, c^{-1}, d^{-1}$) /* Eq. 5 (main draft) */
  return reconstructed
end function



(a) 1% Pixel Budget (c) 3% Pixel Budget

Figure 3. Question-category-wise performance on A-OKVQA dataset [8] at a 1% and 3% pixel budget with Qwen2.5-VL.

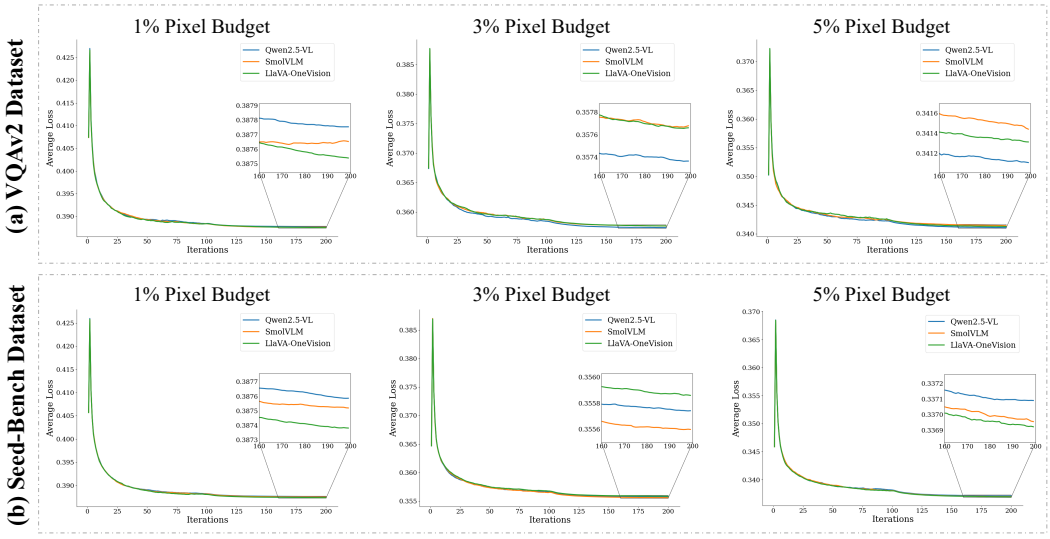


Figure 4. **Loss curves** of Qwen2.5-VL, SmolVLM, and LLaVA-OneVision on **(a)** VQAv2 and **(b)** SEED-Bench datasets under 1%, 3%, and 5% pixel budgets. Insets highlight late-stage convergence differences across models and budgets. **Zoom in for a better view**.

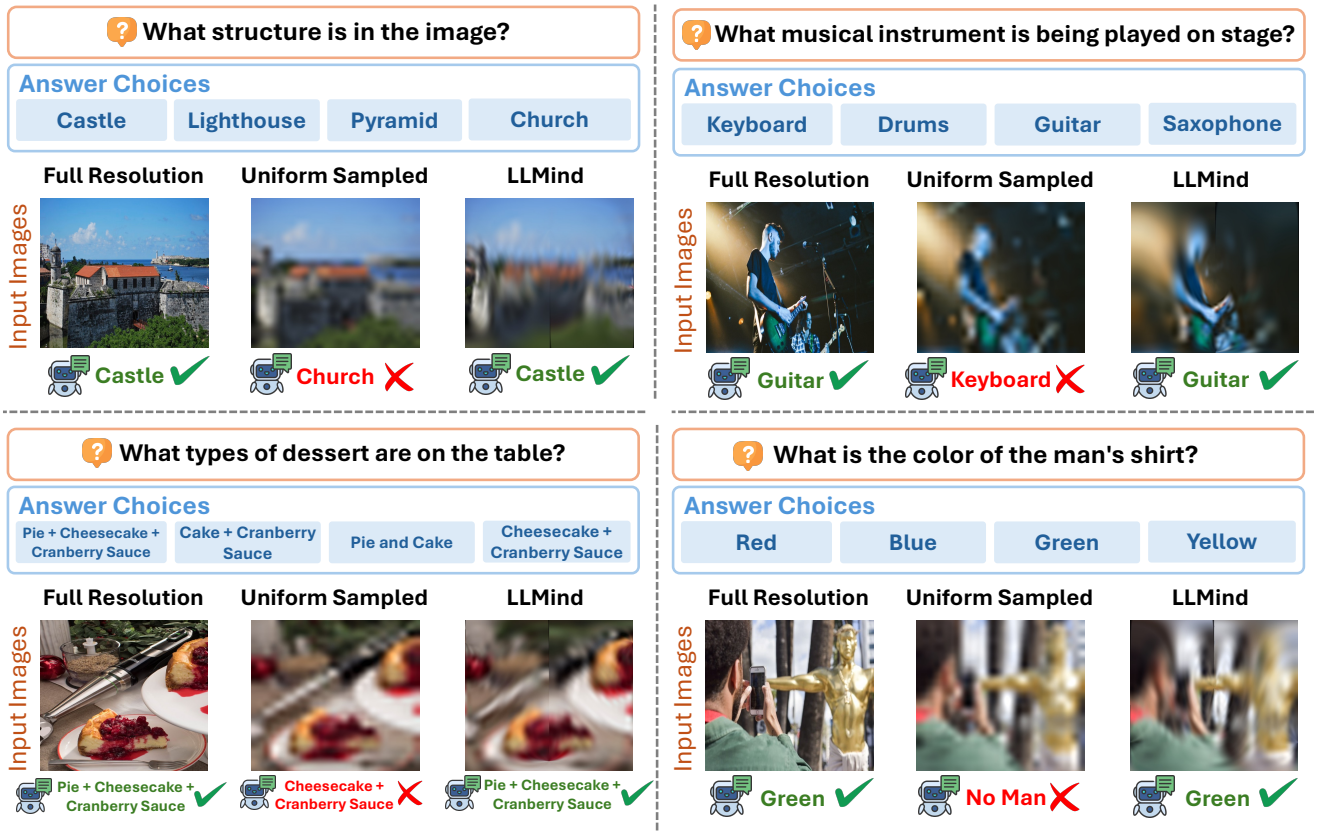


Figure 5. **Qualitative comparison on Seed-Bench [4] dataset with Gemini-2.5-Flash [2] (*black-box*) under 5% pixel budget**. LLMind adaptively allocates resolution to semantically important regions, preserving visual evidence critical for answering the question.

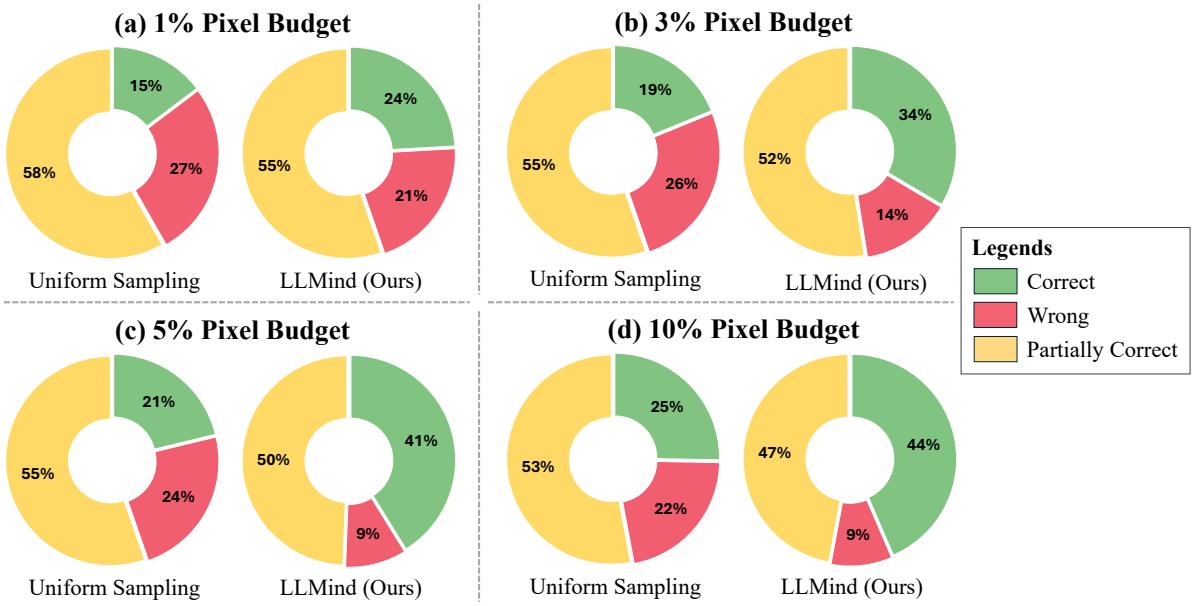


Figure 6. Comparison of prediction outcomes (correct, partially correct, and wrong) for Uniform Sampling and LLMind across four pixel budgets (1%, 3%, 5%, and 10%) on multiple region-specific classification (RSC) tasks. LLMind consistently increases the proportion of correct responses while reducing errors, with improvements becoming more pronounced at higher pixel budgets.

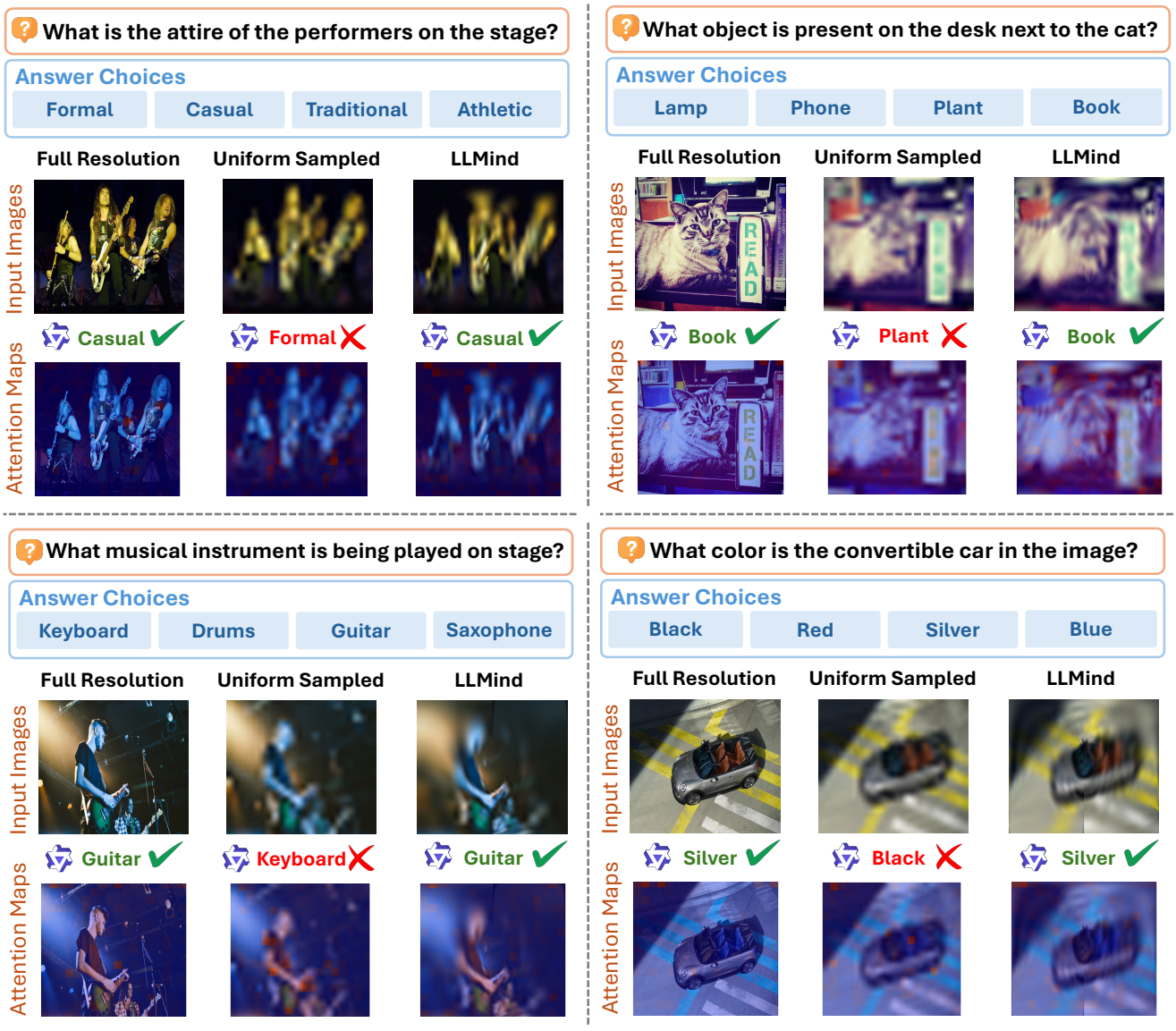


Figure 7. Additional qualitative comparison on Seed-Bench [4] dataset with Qwen2.5-VL under 5% pixel budget.

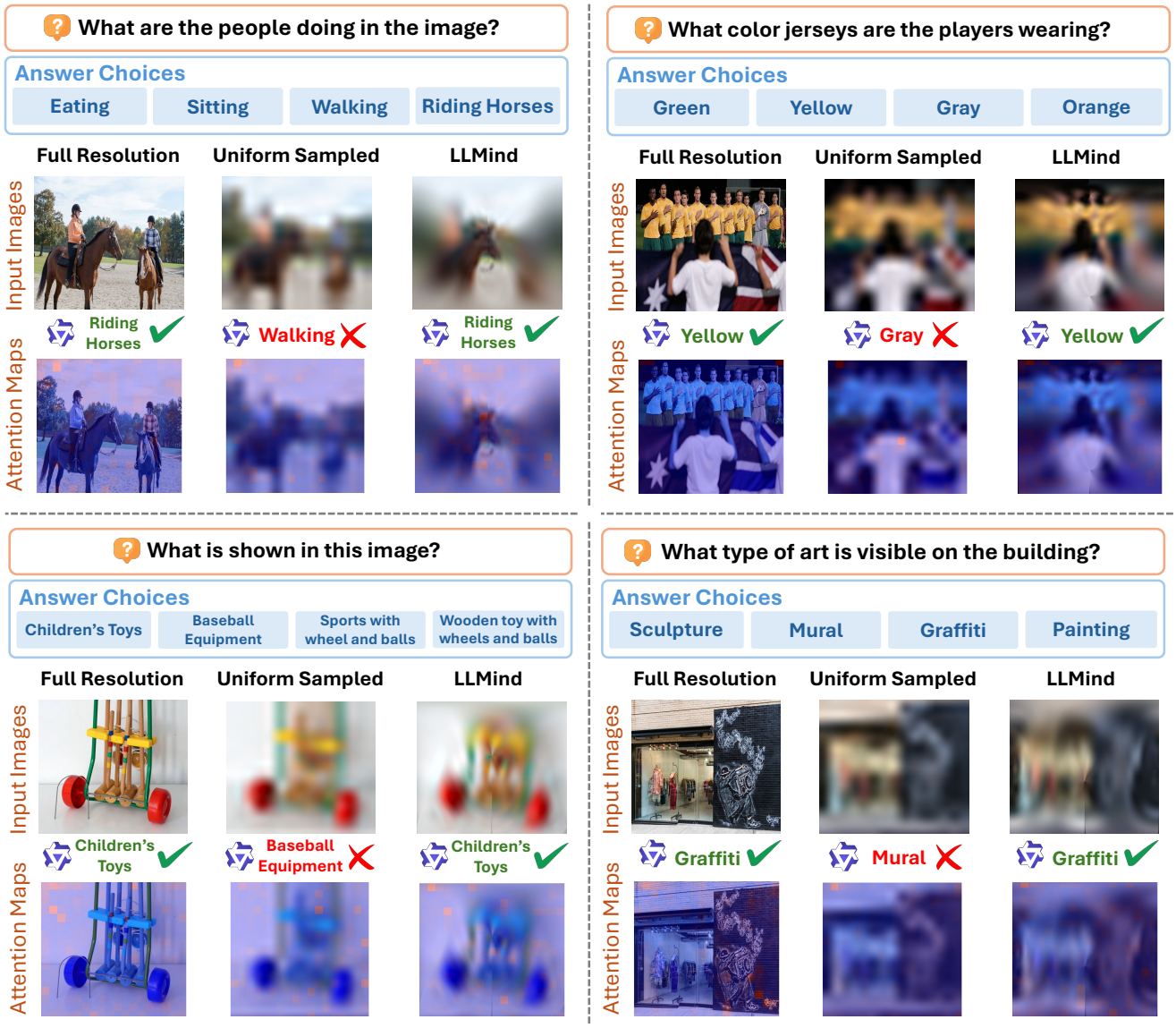


Figure 8. Qualitative comparison on Seed-Bench [4] dataset with Qwen2.5-VL under 3% pixel budget.