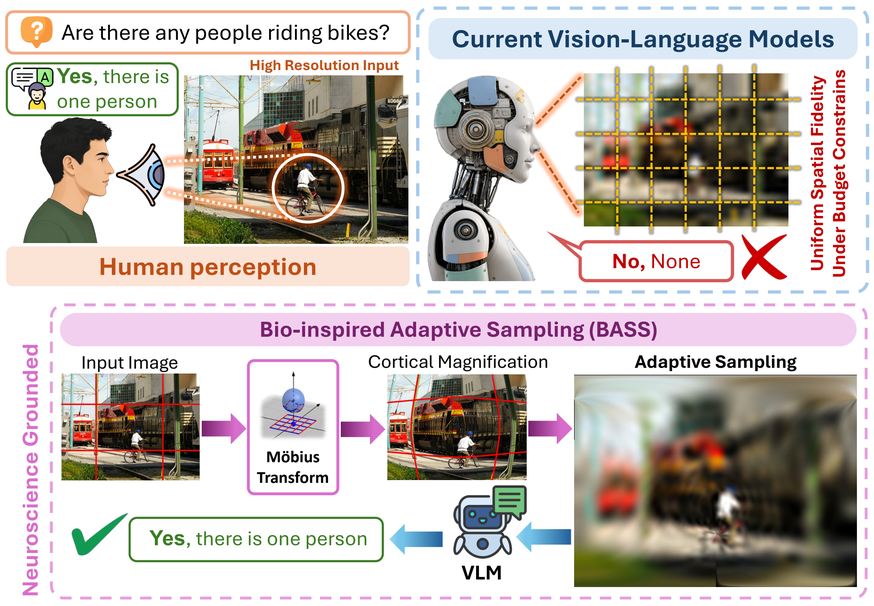
LLMind: Bio-inspired Training-free Adaptive Visual Representations for Vision-Language Models

Soumyaratna Debnath, Bui Duc Manh, Zinan Liu, Lin Wang*
EmPACT Lab, Nanyang Technological University, Singapore

CVPR JUNE 3-7, 2026 DENVER COLORADO
CVPR 2026 Highlight

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation



Current VLMs assume uniform spatial fidelity, allocating equal computation to all pixels.
This leads to **wasted resources on uninformative regions** and loss of critical details.

In contrast, human vision is inherently **adaptive**, **selective**, **efficient**, and **task-driven**, guided by cortical magnification and saccadic eye movements toward task-relevant regions.

Research Question.

Can biologically inspired sampling strategies enable VLMs to achieve higher reasoning efficiency and accuracy than conventional uniform sampling under limited pixel budgets?

Contributions

01 A bio-inspired training-free adaptive sampling framework for efficient VLM perception.

02 A Möbius transformation-based sampling strategy for non-uniform pixel allocation.

03 A model agnostic, closed-loop semantic feedback (CSF) for test-time alignment.

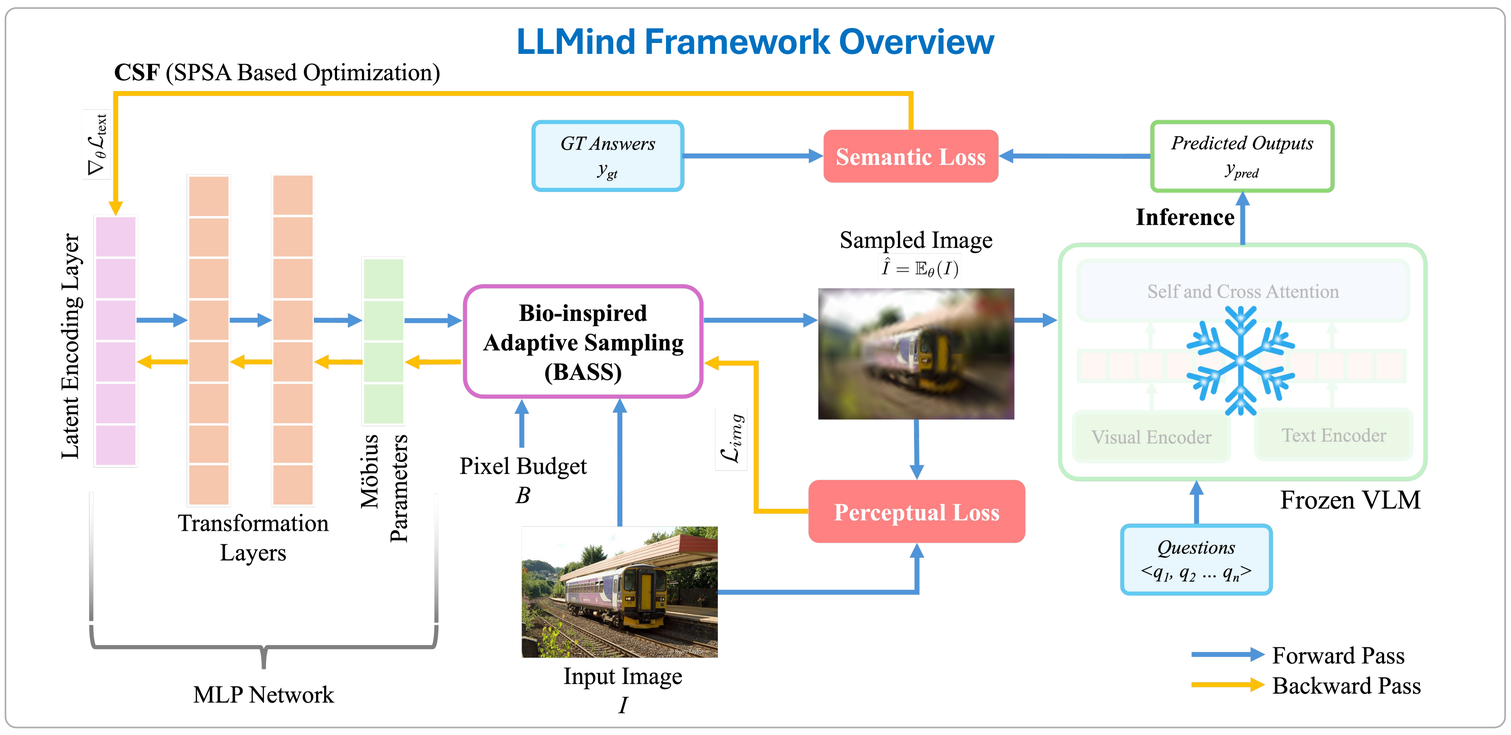
04 Consistent VQA gains under extreme pixel budget constrains across diverse settings.

References

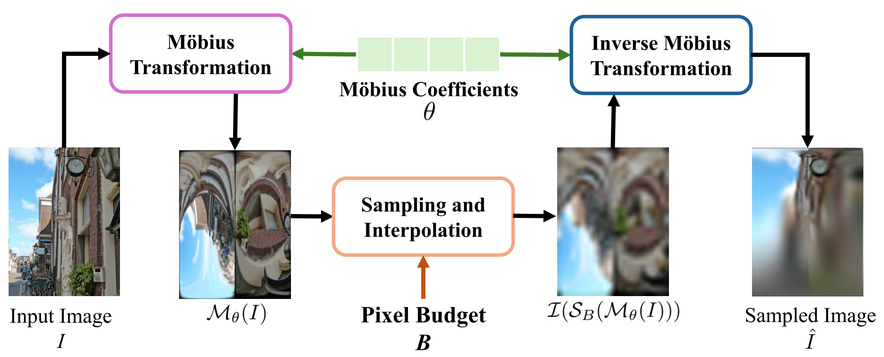
[1] Gizdov, Andrey, Shimon Ullman, and Daniel Harari. "Seeing more with less: human-like representations in vision models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
[2] Killick, George, et al. "Foveation in the Era of Deep Learning." (2023).
[3] Robert-Inacio, Frédérique, et al. "Biologically inspired image sampling for electronic eye." 2010 Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2010.
[4] Bai, Shuai, et al. "Qwen3-vl technical report." (2025).
[5] Zhou, Sharon, et al. "Data augmentation with Mobius transformations." ML: Science and Technology (2021º.

Methodology

LLMind Framework Overview



Bio-inspired Adaptive Sampling Strategy (BASS)



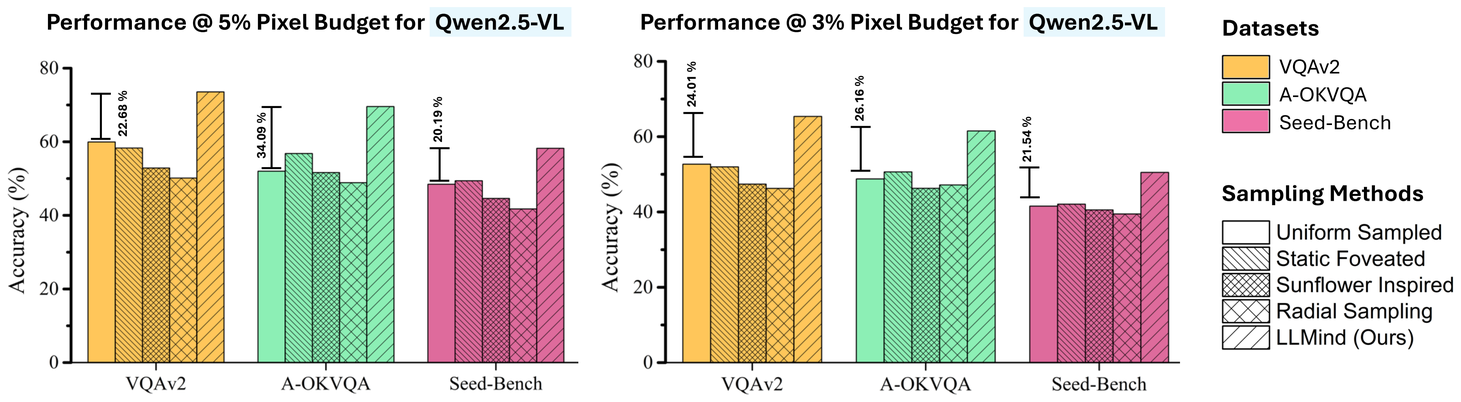
BASS models **cortical magnification** by warping images toward task-relevant regions, while CSF mimics **saccadic eye movements** through iterative refinements.

Non-Uniform Sampling

Experiments and Results

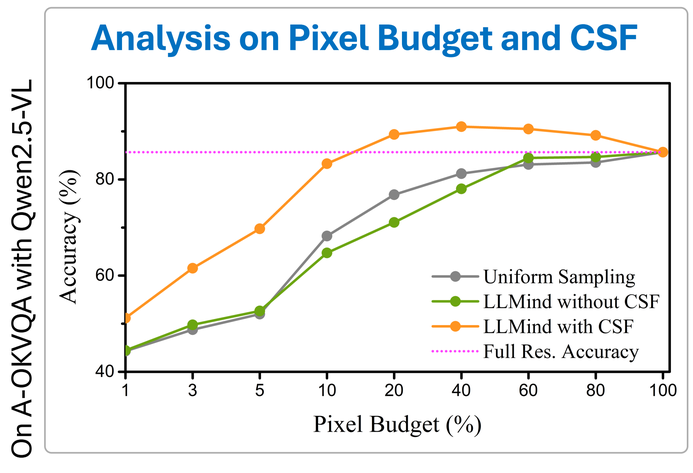
Q-1: *Does LLMind improve scene-level VQA accuracy across diverse benchmarks ? [**Yes**]*
Q-2: *Can LLMind enhance region-guided VQA by reducing background distraction ? [**Yes**]*
Q-3: *What is the impact of the CSF on LLMind ? [**CSF Improves Reasoning in VLMs**]*

Quantitative Results for LLMind Across Various VQA Datasets

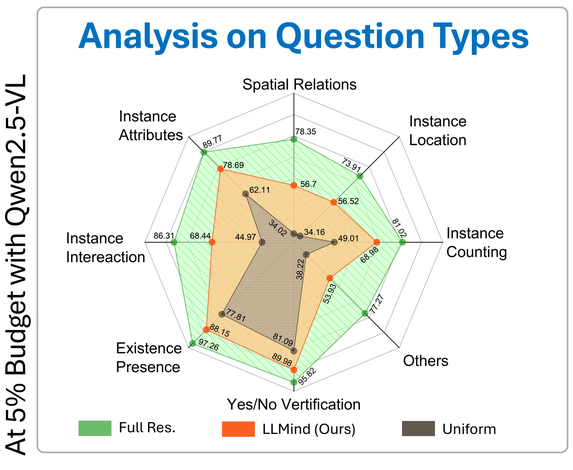


Qualitative Results for LLMind using Qwen2.5-VL at 5% Pixel Budget

Analysis on Pixel Budget and CSF



Analysis on Question Types



Project Page
EmPACT Lab