BEYOND ATTENTION BLACK HOLES: ENERGY LANDSCAPE GUIDANCE FOR RECTIFYING FEATURE STAGNATION IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models have achieved remarkable success in multimodal reasoning, yet they remain surprisingly vulnerable to low-level visual distraction. In this work, we identify and formalize a ubiquitous but overlooked phenomenon: Attentional Drift induced by artificial image padding that corresponds to redundant regions in natural scenes. We observe that semantically vacant black pixels and low-entropy background patches often act as attentional black holes, disproportionately absorbing the model's focus and leading to severe reasoning degradation. Our investigation reveals that this drift is triggered by a Dual Trap inherent in the Transformer architecture: 1) a positional bias toward the end of the sequence, and 2) an attention sink induced by feature stagnation, where stable and uninformative tokens act as attention sinks. To mitigate this universal issue, we propose Energy Landscape Guidance (ELG), a training-free, plug-and-play framework that conceptualizes the evolution of hidden states as trajectories within a high-dimensional manifold. ELG introduces an Energy Scoring Module that leverages spectral analysis to capture evolution flux, thereby isolating various forms of pathological tokens without the need for prior masks. This is followed by a Latent Rectification Module that steers hidden states toward high-information semantic basins via gradient-based intervention during inference. Extensive evaluations on VMCBench demonstrate that ELG consistently rehabilitates various VLMs from visual illusions. Notably, Qwen2.5-VL-7B equipped with ELG outperforms 72B-scale competitors and proprietary models like GPT-4o.

1 INTRODUCTION

Large Vision-Language Models have demonstrated remarkable capabilities in multimodal reasoning and instruction following. However, despite their sophisticated architectures, they remain surprisingly vulnerable to low-level 'visual illusions' (Darcet et al., 2023; Xiao et al., 2023; Li et al., 2024; Liu et al., 2024). A ubiquitous yet underexplored phenomenon occurs during image preprocessing: to align aspect ratios, images are frequently padded with black borders. Through preliminary experiments, we observe a counter-intuitive phenomenon where these semantically devoid, pure-black pixels act as 'black holes,' disproportionately absorbing the model's attention (Sun et al., 2024; Kazemnejad et al., 2023; Bondarenko et al., 2021). This results in a catastrophic attention shift, where VLMs fail to focus on the central informative objects, leading to severe hallucination and reasoning degradation.

What makes these meaningless black pixels so disruptive to the powerful self-attention mechanism? Through a deep inspection of the hidden states in VLMs, as illustrated in Fig.1, we reveal that letterboxing inadvertently triggers a Dual Trap deeply rooted in the Transformer architecture. ❶ The first mechanism is the End-of-Sequence Position Bias. During the raster-scan tokenization of images, the bottom black borders inevitably fall at the very end of the sequence. Due to the inherent nature of autoregressive training and artifacts in positional encodings, VLMs naturally exhibit a recency bias, erroneously allocating excessive attention weights to these trailing tokens. ❷ The second, and more insidious mechanism, is the *Feature Stagnation-induced Attention Sink*. Specifically, we pad the letterboxing to surround the image. Since the padding pixels are completely black (zero information), their corresponding hidden states remain highly homogeneous and essentially stagnant across Transformer layers (i.e., $\Delta h \approx 0$). The Attention Sink phenomenon can be viewed as a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

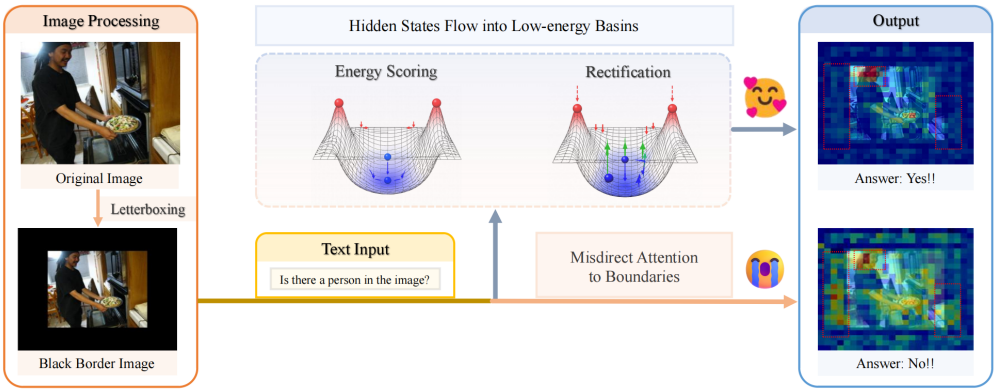


Figure 1: Overview of ELG for mitigating attention drift. (Left) Padded images often trigger spurious attention on boundaries, causing incorrect VLM outputs. (Middle) ELG rectifies latent trajectories by modeling an energy landscape where hidden states are pushed toward low-energy semantic basins. (right) While the original attention map attends to the black borders, the rectified version successfully re-localizes attention to the foreground. This training-free intervention successfully refocuses attention on salient objects and corrects model reasoning.

computational shortcut—*a form of 'laziness'* in essence—whereby the Transformer redirects excess attention mass toward tokens with stable but semantically empty representations, so as to preserve the numerical stability of Softmax. Consequently, the surrounding black borders form a 'siege', silently draining attention away from the high-information central objects.

We posit that letterboxing is merely the most conspicuous and extreme manifestation of the above fundamental vulnerability. The 'Dual Trap' is *not confined to artificial padding around the image*; rather, it is typical of the semantically sparse or redundant regions within the original images. For example, in real-world scenarios, uninformative elements such as out-of-focus backgrounds, clear skies, or repetitive textures frequently exhibit similar feature stagnation characteristics (Bolya and Hoffman, 2023; Huang et al., 2024). Similar to the all-zero letterbox, these regions act as latent attractors that silently divert the model's focus away from salient content (Liu et al., 2024). This suggests that the attentional drift observed in padded images is, in fact, a common issue in VLMs for the vast majority of images. Therefore, ***the scope of this work is not confined to padded images***. Instead, we present a ***general-purpose method designed to address this issue across arbitrary input images***.

To mitigate this distortion without the need for expensive retraining or fine-tuning, we propose Energy Landscape Guidance (ELG), a training-free framework that conceptualizes the evolution of hidden states as trajectories within a high-dimensional manifold. ELG reshapes the latent space into a content-aware energy landscape where 'pathological' tokens are assigned high energy. Specifically, we introduce an Energy Scoring Module that integrates dynamic evolution flux, boundary-aware geometric priors, and semantic consistency to identify uninformative regions. Subsequently, a Latent Rectification Module employs gradient-based intervention during inference to steer hidden states toward high-information 'semantic valleys'.

Our framework offers several distinct advantages. First, it is plug-and-play, requiring no architectural changes or parameter updates. Second, it is content-aware, dynamically adapting the rectification intensity based on the informational value of each token. We evaluate ELG on the VMCBench benchmark, comprising 9,450 samples across diverse VQA tasks. Experimental results show that ELG significantly improves the performance of various VLMs, with our 7B-scale model even outperforming 72B-scale competitors and proprietary models like GPT-4o in specific reasoning tasks.

The core contributions of this work are summarized as follows:

- We identify and formalize the attentional drift phenomenon in VLMs through intentional image padding, revealing a Dual Trap involving position bias and feature stagnation.
- We propose Energy Landscape Guidance, a novel training-free inference-time framework that rectifies hidden states by navigating a composite energy landscape.

- We introduce the Energy Scoring Module, which utilizes spectral analysis and geometric priors to robustly detect and suppress uninformative 'pathological' tokens.
- Extensive experiments demonstrate that ELG yields consistent improvements across multiple VLMs architectures, achieving a new state-of-the-art on the VMCBench benchmark and effectively rehabilitating models from visual illusions.

2 Related Work

2.1 Inference-time Intervention and Visual Guidance

The evolution of Vision-Language Models has increasingly pivoted toward enhancing complex reasoning through dynamic state modulation. While conventional Adaptation methods, such as Visual Prompt Tuning (Jia et al., 2022) or Parameter-Efficient Fine-Tuning (Hu et al., 2022), seek to align modal representations via learnable tokens, they remain largely static post-training and lack the corrective granularity to mitigate low-level perceptual artifacts. Concurrently, the emerging paradigm of Inference-time Computing exemplified by test-time scaling (Zhang et al., 2025a; Muennighoff et al., 2025) and trajectory search (Yao et al., 2023) aims to refine model outputs without gradient-based parameter updates (Chen et al., 2018; Bronstein et al., 2021). However, these methods primarily operate within high-level semantic spaces, often ignoring the underlying manifold geometry where "attention sinks" and spatial distortions originate. Departing from these abstract updates, our work introduces a Geometrically Grounded Guidance strategy (LeCun et al., 2006). By formalizing the hidden state evolution as a continuous-time trajectory within a high-dimensional energy landscape, we inject explicit spatial-semantic constraints to rectify latent representations. This approach bridges the gap between inference-time search and geometric inductive biases, enabling a more robust and interpretable rectification of spurious attention patterns.

2.2 Improving Visual Attention and Grounding in VLMs

The landscape of Vision-Language Models has seen significant progress in methods aimed at mitigating attention drift and improving fine-grained visual grounding. Conscious Gaze (Bu et al., 2025) introduces an adaptive attention mechanism that detects moments when visual grounding is required and reorients mid-layer attention toward visual tokens at inference time. This approach leads to reduced hallucination and improved grounding performance across various VLMs architectures. Direct Visual Grounding (Rasheed et al., 2024) proposes a supervision mechanism that directly constrains visual token attention through a KL-divergence based loss toward ground-truth attention maps, enabling better localization of visual features relevant to queries. Attention Guided Efficient VLMs (Sha et al., 2026; Chen et al., 2023; You et al., 2023; Kang et al., 2025; Choi et al., 2026; Zhang et al., 2026; Guan et al., 2024) leverage segmentation knowledge extracted from auxiliary models to guide cross-attention, improving semantic alignment between the vision and language modalities. Region-aware Vision Language Models further address localization accuracy by incorporating region proposal mechanisms and IoU-aware objectives, enhancing the focus on relevant regions in complex images. Additionally, iterative prompt-based feedback (Yin et al., 2024) frameworks have been shown to improve semantic grounding by utilizing binary verification signals to refine VLMs predictions during inference. In contrast, our method integrates an energy-based rectification strategy and latent representation adjustment to directly model attention drift and enhance visual focus in VLMs. This approach yields consistent improvements over prior methods, providing a more direct and robust solution for attention correction and visual grounding.

3 Proposed Method

We conceptualize the evolution of hidden states in Vision-Language Models as trajectories within a high-dimensional manifold (Zhou et al., 2021; Ramsauer et al., 2020; Hopfield, 1982; Grathwohl et al., 2019). Let $\mathcal{H} = \{\mathbf{H}^1, \ldots, \mathbf{H}^L\}$ denote the sequence of hidden state tensors across L Transformer layers, where $\mathbf{H}^l \in \mathbb{R}^{B\times T\times D}$. The core problem of Attention Drift is thus defined as the latent states of uninformative tokens acting as spurious attractors that disproportionately dominate the self-attention mechanism.

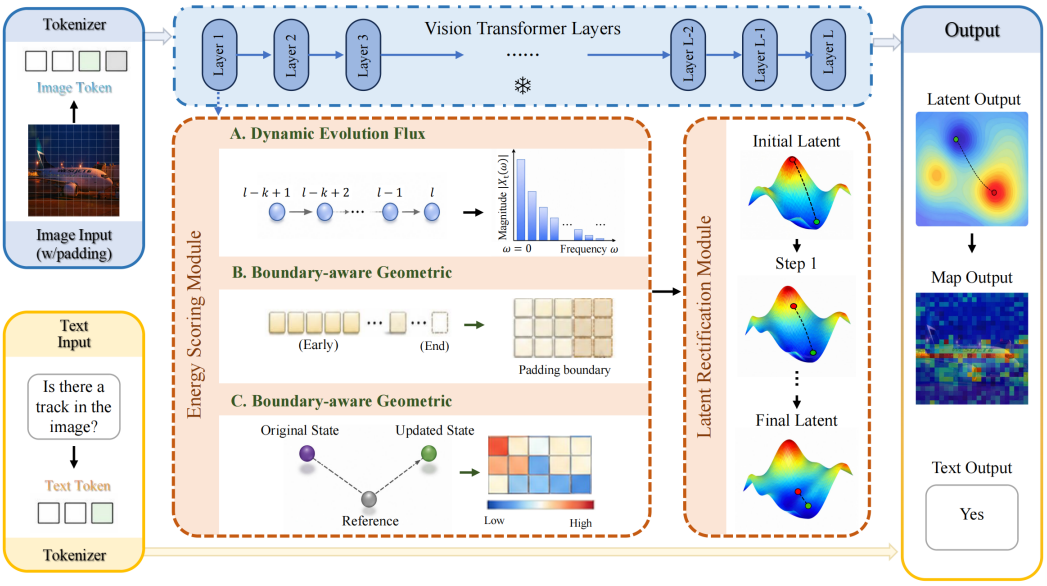


Figure 2: The framework of Energy Landscape Guidance (ELG). ELG employs an Energy Scoring Module to detect uninformative regions and a Boundary-aware Geometric approach to adjust uninformative impacts. The Latent Rectification Module then refines the hidden states, steering them towards informative regions, resulting in improved reasoning and accurate outputs.

To address this challenge, as illustrated in Fig. 2, we introduce Energy Landscape Guidance (ELG), a training-free framework designed to reshape the latent space into a content-aware energy landscape (we provide the theoretical foundation of the energy landscape in the Appendix A). We decompose ELG into two coupled components:

1. **Energy Scoring Module (ESM):** A composite potential function $\mathcal{E}(\mathbf{H}^l)$ that identifies pathological tokens across both spectral and spatial domains.
2. **Latent Rectification Module (LRM):** An inference-time engine that employs gradient-based intervention to steer hidden states toward high-information 'valleys'.

3.1 Energy Scoring Module

The Energy Scoring Module (ESM) establishes a content-aware objective for hidden-state rectification by identifying *pathological tokens* whose representations exhibit abnormal stagnation or positional vulnerability. Such tokens, often corresponding to padding artifacts or low-information regions, may attract disproportionate attention and distort the latent representation dynamics.

To characterize these effects, we construct a composite pathological energy, $\mathcal{E}$, for each token t at layer l. It integrates three complementary factors: **stagnation energy** ($\mathcal{E}_{\text{stag}}$), derived from dynamic evolution flux; a **boundary-aware geometric prior** (Ψ_{geo}); and **semantic consistency** (Ω_{cons}). The resulting objective is

$$\mathcal{E}(\mathbf{h}_t^l) = \alpha \mathcal{E}_{\text{stag}}(\mathbf{h}_t^l) + \beta \Psi_{\text{geo}}(\mathbf{h}_t^l, t) + \gamma \Omega_{\text{cons}}(\mathbf{h}_t^l), \tag{1}$$

where α, β, and γ control the relative contributions of the three terms. This formulation assigns higher energy to stagnant and positionally vulnerable representations while constraining rectification to preserve the original semantics. Consequently, minimizing $\mathcal{E}$ provides a consistent objective for steering pathological hidden states toward more informative latent dynamics.

3.1.1 Dynamic Evolution Flux

The stagnation energy is motivated by the observation that uninformative tokens often exhibit weak representational evolution across Transformer layers. With residual connections, the hidden state of

token t evolves as

$$\mathbf{h}_t^l = \mathbf{h}_t^{l-1} + \text{SubLayer}(\mathbf{h}_t^{l-1}), \tag{2}$$

where SubLayer$(\cdot)$ denotes the self-attention and feed-forward operations within a Transformer block. Informative tokens typically undergo persistent semantic refinement, whereas uninformative tokens, such as padding artifacts, may rapidly approach a nearly stationary state with negligible layer-wise updates.

To quantify this behavior, we treat the recent hidden states $\mathcal{T}_t = \{\mathbf{h}_t^{l-k}\}_{k=0}^{K-1}$ as a discrete trajectory and apply the Discrete Fourier Transform (DFT):

$$\mathbf{X}_t[\omega] = \sum_{k=0}^{K-1} \mathbf{h}_t^{l-k} \exp\left(-j\frac{2\pi}{K}\omega k\right). \tag{3}$$

The zero-frequency component captures the stationary part of the trajectory, while non-zero frequencies characterize layer-wise representation changes. We therefore define the evolution flux as

$$\Phi_{\text{flux}}(\mathbf{h}_t^l) = \log\left(1 + \frac{1}{K}\sum_{\omega=1}^{K-1} \|\mathbf{X}_t[\omega]\|_2^2\right). \tag{4}$$

A high evolution flux indicates persistent representation refinement, whereas a low value indicates latent stagnation. Since Φ_{flux} is a refinement score rather than a pathological energy, we convert it into a minimizable stagnation term:

$$\mathcal{E}_{\text{stag}}(\mathbf{h}_t^l) = -\log\left(\Phi_{\text{flux}}(\mathbf{h}_t^l) + \epsilon\right), \tag{5}$$

where $\epsilon > 0$ ensures numerical stability. Thus, stagnant tokens receive high energy, while actively evolving representations incur a smaller penalty. Minimizing $\mathcal{E}_{\text{stag}}$ therefore encourages hidden states to move away from representational stagnation.

3.1.2 Boundary-aware Geometric Prior

Evolution flux alone does not determine whether a stagnant token is harmful. Its effect also depends on its position in the autoregressive sequence. In particular, tokens appearing later in the visual sequence can be more susceptible to recency bias. When low-flux tokens near image boundaries are mapped toward the tail of a rasterized sequence, they may consequently behave as spurious "Attention Sinks."

To jointly model stagnation and positional vulnerability, we replace hard token membership with a differentiable stagnation gate:

$$s_t^l = \sigma\left(\frac{\tau - \Phi_{\text{flux}}(\mathbf{h}_t^l)}{\delta}\right), \tag{6}$$

where τ is an adaptive threshold estimated from the flux distribution, $\delta > 0$ controls the transition smoothness, and $\sigma(\cdot)$ denotes the sigmoid function. The gate approaches one for low-flux tokens and decreases as their representations become more dynamically active.

We then define the Boundary-aware Geometric Prior as

$$\Psi_{\text{geo}}(\mathbf{h}_t^l, t) = s_t^l \cdot \ln\left(1 + \text{pos}(t)\right), \tag{7}$$

where $\text{pos}(t)$ denotes the relative sequential position of token t. This formulation assigns larger penalties to stagnant tokens occurring at positionally vulnerable locations. Importantly, because s_t^l depends continuously on Φ_{flux}, the geometric prior remains differentiable with respect to the hidden state and can directly participate in latent rectification. The logarithmic position term prevents excessive amplification while retaining stronger guidance for later tokens.

3.1.3 Semantic Consistency

Directly modifying hidden states may distort the pretrained representation manifold and compromise semantic information. We therefore introduce a directional consistency regularizer that constrains rectification to remain close to the original representation.

Let $\hat{\mathbf{h}}_t^l$ denote the hidden state before rectification. We define

$$\Omega_{\text{cons}}(\mathbf{h}_t^l) = 1 - \frac{\mathbf{h}_t^l \cdot \hat{\mathbf{h}}_t^l}{\|\mathbf{h}_t^l\|_2 \|\hat{\mathbf{h}}_t^l\|_2}. \tag{8}$$

The cosine-distance penalty discourages large angular deviations from the original hidden state while allowing its magnitude and local trajectory to be adjusted. It therefore acts as a semantic anchor during inference-time intervention, preventing the rectified representation from drifting excessively from the pretrained latent space.

Together, the three terms define a coherent rectification objective: $\mathcal{E}_{\text{stag}}$ penalizes insufficient representation evolution, Ψ_{geo} strengthens the penalty for stagnant tokens at vulnerable sequential positions, and Ω_{cons} limits semantic distortion. The resulting energy can therefore be minimized directly by the Latent Rectification Module to suppress pathological stagnation while preserving the pretrained representation structure.

3.2 Latent Rectification Module

The Latent Rectification Module (LRM) steers the evolution of representations by refining the output of each Transformer layer. For each layer l, let $f^{(l)}(\cdot)$ denote the standard forward transformation. The rectification process is executed as a single-step optimization following the vanilla computation. Given the rectified input $\mathbf{H}^{l-1*}$ from the preceding layer, the transition to the subsequent state can be expressed as:

$$\mathbf{h}_t^l = f^{(l)}(\mathbf{H}^{l-1*})_t, \tag{9}$$

Then, the hidden state $\mathbf{h}_t^l$ is rectified to obtain $\mathbf{h}_t^{l*}$:

$$\mathbf{h}_t^{l*} = \mathbf{h}_t^l - \eta \cdot \frac{\partial \mathcal{E}(\mathbf{h}_t^l)}{\partial \mathbf{h}_t^l}, \tag{10}$$

where $\mathbf{h}_t^l$ represents the vanilla hidden state, and $\mathbf{h}_t^{l*}$ denotes its rectified counterpart. The term $\frac{\partial \mathcal{E}}{\partial \mathbf{h}_t^l}$ represents the gradient of the loss function with respect to the hidden state, which indicates how the current state should be adjusted to avoid representational stagnation and geometric bias. These biases could lead to instability or inconsistency in the model's performance. The purpose of the rectification is to encourage more diverse and effective evolution of the representations. The coupling coefficient η controls the intensity of the rectification, ensuring that the process remains a local perturbation without causing disruptive changes to the model's behavior, thereby maintaining its stability.

4 Experiments

Datasets and Evaluation Protocol. We evaluate the ELG framework on VMCBench (Zhang et al., 2025b), a comprehensive benchmark comprising **9,450** multiple-choice samples across 20 diverse VQA datasets. This benchmark covers four core dimensions: General Scene Understanding, Complex Reasoning, Text-Rich & OCR, and Structured Document Understanding. We adopt Accuracy as the primary metric to ensure evaluation stability.

Table 1: Pilot study on the impact of black padding. Accuracy (%) is averaged over the spatial-reasoning tasks in the benchmark. Δ_{pad} denotes the performance drop caused by letterboxing relative to the original input, while the values in parentheses indicate the improvement of ELG over the corresponding letterboxed input.

Model	Original	Original + ELG	Letterboxed	Δ_{pad}	Letterboxed + ELG (Ours)
LLaVA-1.5-7B	58.2	64.0	42.4	−15.8	54.3 (+11.9)
Phi-3-Vision	70.3	80.4	58.6	−11.7	68.7 (+10.1)
Qwen2-VL-7B	80.9	87.1	69.3	−11.6	80.2 (+10.9)
Qwen2.5-VL-7B	83.9	88.6	73.2	−10.7	**84.9** (+11.7)

Pilot Study. To empirically verify the deleterious impact of padding artifacts, we conduct a pilot study across various VLM architectures by comparing their performance on original versus letterboxed inputs. As summarized in Table 1, the introduction of semantically vacant black borders triggers a consistent performance collapse across all models, with accuracy drops (Δ) ranging from **10.7%** to **15.8%**. This significant degradation provides strong empirical evidence for our hypothesis: uninformative padding regions act as 'spurious attractors' that drain attention from salient objects. Notably, our proposed ELG framework not only successfully mitigates these negative effects but effectively rehabilitates the models' reasoning capabilities. For the Qwen2.5-VL-7B backbone, ELG elevates the accuracy from a degraded **73.2%** to **84.9%**, representing a substantial **+11.7%** improvement over the letterboxed baseline.

Comparison with State-of-the-arts. Table 2 provides a comprehensive comparison of our proposed ELG framework against various state-of-the-art baselines on the VMCBench. Our method, when applied to the Qwen2.5-VL-7B backbone, achieves a remarkable average score of **88.6%**, yielding an overall improvement of 4.7 percentage points over the base model. This gain is consistent across all evaluation dimensions, with particularly notable boosts in complex categories such as Reasoning (+5.9 percentage points) and Doc & Chart (+6.4 percentage points), where our model achieves **74.1%** and **92.9%** respectively. Furthermore, ELG sets a new state-of-the-art in OCR performance with a score of **97.2%**. Remarkably, our 7B-scale model even outperforms much larger competitors, including the 72B-parameter Qwen2-VL (87.0% Avg.) and leading proprietary models such as GPT-4o (82.9% Avg.) and Claude-3.5-Sonnet (80.5% Avg.). These results demonstrate that the ELG framework significantly enhances the perceptual and reasoning capabilities of vision-language models, achieving superior efficiency by surpassing the performance of models with nearly ten times the parameter count.

Table 2: Performance Comparison on VMCBench. **ELG** denotes our framework applied to *Qwen2.5-VL-7B*. Results are reported as mean$_{\pm \text{std}}$ over three independent runs. Numbers in blue indicate the improvement over the base model (%).

Model	General	Reasoning	OCR	Doc & Chart	Avg.
Proprietary / Large-Scale Models (>30B)					
Qwen2-VL-72B	$88.50_{\pm 0.21}$	$72.60_{\pm 0.32}$	$96.80_{\pm 0.18}$	$90.10_{\pm 0.31}$	$87.00_{\pm 0.22}$
GPT-4o	$85.20_{\pm 0.12}$	$66.90_{\pm 0.28}$	$96.40_{\pm 0.11}$	$83.10_{\pm 0.24}$	$82.90_{\pm 0.19}$
Claude-3.5-Sonnet	$81.30_{\pm 0.23}$	$62.80_{\pm 0.35}$	$93.40_{\pm 0.21}$	$84.60_{\pm 0.33}$	$80.53_{\pm 0.26}$
Molmo-72B	$82.90_{\pm 0.27}$	$66.60_{\pm 0.41}$	$94.70_{\pm 0.23}$	$81.10_{\pm 0.36}$	$81.33_{\pm 0.29}$
Open-Source Mid-Scale Models (10B–40B)					
Qwen2.5-VL-14B	$89.80_{\pm 0.31}$	$72.60_{\pm 0.42}$	$96.50_{\pm 0.25}$	$89.90_{\pm 0.44}$	$87.20_{\pm 0.34}$
Cambrian-34B	$83.70_{\pm 0.42}$	$65.90_{\pm 0.51}$	$95.70_{\pm 0.33}$	$73.30_{\pm 0.56}$	$79.65_{\pm 0.44}$
VILA1.5-40B	$82.50_{\pm 0.45}$	$65.30_{\pm 0.62}$	$93.20_{\pm 0.38}$	$67.40_{\pm 0.65}$	$77.10_{\pm 0.51}$
CogVLMs2-19B	$78.10_{\pm 0.53}$	$55.60_{\pm 0.74}$	$92.30_{\pm 0.42}$	$72.60_{\pm 0.68}$	$74.65_{\pm 0.59}$
Efficient Models ($\leq$8B)					
Qwen3-VL-8B	$87.30_{\pm 0.38}$	$71.80_{\pm 0.49}$	$96.40_{\pm 0.29}$	$89.50_{\pm 0.43}$	$86.25_{\pm 0.40}$
Qwen2.5-VL-7B (Base)	$85.60_{\pm 0.65}$	$68.20_{\pm 0.82}$	$95.30_{\pm 0.52}$	$86.50_{\pm 0.74}$	$83.90_{\pm 0.68}$
Qwen2-VL-7B	$84.50_{\pm 0.72}$	$62.70_{\pm 0.91}$	$96.40_{\pm 0.48}$	$80.10_{\pm 0.83}$	$80.93_{\pm 0.74}$
Molmo-7B-D	$73.20_{\pm 0.61}$	$55.50_{\pm 0.85}$	$91.70_{\pm 0.53}$	$72.10_{\pm 0.77}$	$73.13_{\pm 0.66}$
Cambrian-8B	$77.90_{\pm 0.68}$	$56.40_{\pm 0.88}$	$91.00_{\pm 0.62}$	$65.40_{\pm 0.85}$	$75.18_{\pm 0.71}$
ELG (Ours)	$\mathbf{90.20}_{\pm 0.28}$	$\mathbf{74.10}_{\pm 0.35}$	$\mathbf{97.20}_{\pm 0.18}$	$\mathbf{92.90}_{\pm 0.29}$	$\mathbf{88.60}_{\pm 0.32}$
Improvement	**(+4.60)**	**(+5.90)**	**(+1.90)**	**(+6.40)**	**(+4.70)**

Ablation Study. To dissect the contribution of each component in ELG, we perform an incremental ablation study on VMCBench, as summarized in Table 3. Starting from the Qwen2.5-VL-7B baseline (83.90% Avg.), introducing the flux-based stagnation energy $\mathcal{E}_{\text{stag}}$ improves the average performance to 86.78% (+2.88 points), demonstrating the benefit of penalizing insufficient representation evolution. Incorporating the boundary-aware geometric prior Ψ_{geo} further increases the performance to 88.23% (+4.33 points), suggesting that jointly accounting for latent stagnation and positional vulnerability provides complementary guidance. Finally, adding the semantic consistency regularizer Ω_{cons} yields the full ELG model with an average accuracy of 88.60% (+4.70 points). Although its additional

Table 3: Ablation study of ELG components on VMCBench. We report performance across all sub-categories. The blue values in the Avg. column indicate the cumulative improvement over the Base model (percentage points).

Configuration	$\mathcal{E}_{\text{stag}}$	Ψ_{geo}	Ω_{cons}	Gen.	Reas.	OCR	Doc.	Avg.
Base (Qwen2.5-VL-7B)	–	–	–	85.6	68.2	95.3	86.5	83.90
+ Stagnation Energy	✓	–	–	89.4	72.3	96.5	88.9	86.78 (+2.88)
+ Geometric Prior	✓	✓	–	90.0	73.7	97.8	91.4	88.23 (+4.33)
+ Semantic Cons. (Full)	✓	✓	✓	**90.2**	**74.1**	**97.2**	**92.9**	**88.60** (+4.70)

numerical gain is moderate, the consistency term constrains hidden-state rectification from deviating excessively from the original pretrained representation.

Qualitative Results. We present qualitative results comparing the performance of the Qwen2.5-VL-7B model and our method. The visual attention maps provide insights into the model's ability to focus on relevant objects in the images for various tasks. For each image, we evaluate whether the model correctly identifies the object in question, and compare the attention maps of the baseline model and our method. As shown in Figure 3, the attention maps generated by the Qwen2.5-VL-7B model exhibit some spurious regions, especially in images with ambiguous or distracting backgrounds. Our method, however, demonstrates improved focus on the key objects in the scene, resulting in more accurate predictions. Specifically, in the task of identifying a person, our method highlights the individual more clearly, whereas the baseline model struggles to isolate the relevant area. Similarly, for the tasks involving birds or backpacks, our method more accurately identifies these objects, leading to improved prediction accuracy. Additionally, visualizations under complex scenarios and degraded scenes are provided in the Appendix B.3.

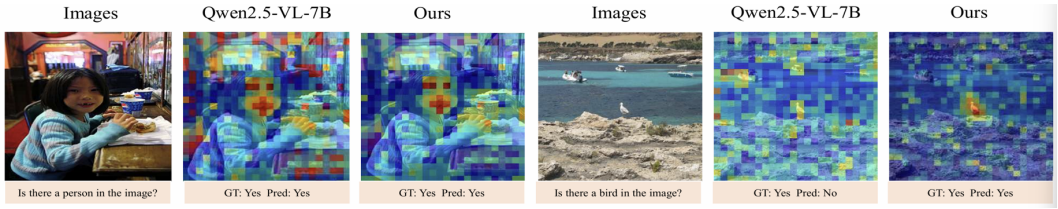


Figure 3: Qualitative comparison of attention maps between the baseline model (Qwen2.5-VL-7B) and our method.

5 Analysis Study

Table 4: Comparison of Attention Drift and Visual Grounding Correction Methods on VMCBench. All methods are evaluated using Qwen2.5-VL-7B as the backbone for a fair comparison. **Ours** refers to the proposed Energy Landscape Guidance (%).

Method	General	Reasoning	OCR	Doc.	Avg.
Qwen2.5-VL-7B (Base)	85.6	68.2	95.3	86.5	83.90
Direct Visual Grounding	88.2	70.3	95.7	87.8	85.50
VAR	88.7	70.8	96.0	88.5	86.00
Conscious Gaze	89.8	72.6	96.5	89.9	87.20
LGS	89.6	72.8	96.4	90.8	87.40
R-VLMs	89.5	73.0	96.2	92.0	87.68
AGE-VLMs	89.4	73.6	96.8	91.4	87.80
ELG (Ours)	**90.2**	**74.1**	**97.2**	**92.9**	**88.60**

Comparison with Attention Correction Methods. To further validate the efficacy of ELG, we compare it against state-of-the-art attention and grounding correction strategies on VMCBench. To ensure a rigorous comparison, we re-implement competitive inference-time methods (e.g., Conscious Gaze (Bu et al., 2025)) and representative architectural solutions on the same Qwen2.5-VL-7B

backbone. As shown in Table 4, ELG achieves a state-of-the-art average score of **88.6%**. Notably, ELG outperforms R-VLMs (You et al., 2023) and AGE-VLMs (Cai et al., 2024) 0.92 and 0.80 percentage points, respectively, despite the latter relying on auxiliary segmentation priors or extensive region-based pre-training. This demonstrates that by directly rectifying the latent energy landscape, ELG can recover fine-grained visual focus more effectively than heuristic-based or supervision-heavy alternatives.

Efficiency Analysis. As shown in Figure 4, efficiency benchmarks demonstrate that ELG consistently occupies the Pareto-optimal frontier across throughput, latency, and GFLOPS. By replacing complex iterative prompting or auxiliary models with a single-step, gradient-based hidden state rectification, ELG achieves a peak throughput of 62 imgs/sec and a minimal latency of 0.016s/sample—representing a 15–25% reduction compared to competitive baselines—while maintaining a superior 88.6% accuracy. This highlights the framework's capacity to deliver robust multimodal reasoning with negligible computational overhead.

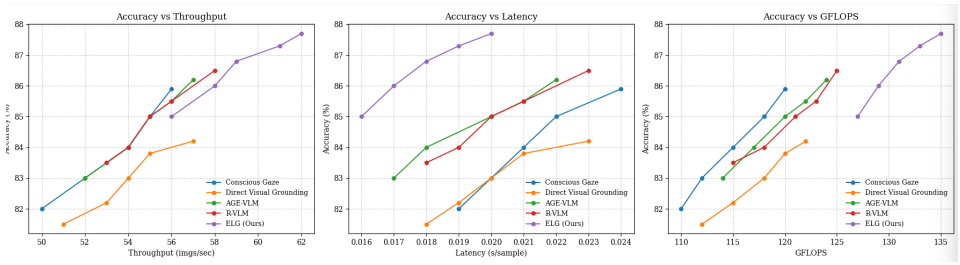


Figure 4: Efficiency evaluation of Energy Landscape Guidance against state-of-the-art attention correction methods on VMCBench.

Generalization Analysis. To evaluate the architectural robustness of the ELG framework, we apply it to various mainstream backbones, including LLaVA-1.5, InternVL-2, and the Qwen series. As summarized in Table 5, ELG consistently yields significant performance gains across all architectures, with an average improvement of **+5.6%**. Notably, the framework achieves a maximum enhancement of **+6.2%** on the Qwen2-VL-7B backbone, while also providing substantial gains for newer models like Qwen3-VL-8B (**+5.3%**). The steady improvements in specialized dimensions, such as Reasoning (**+6.2%**) and Doc&Chart (**+5.7%**), demonstrate that ELG effectively refines latent representations regardless of the underlying model design or scale, confirming its superior generalization capabilities.

Table 5: Cross-architecture generalization. Base Acc. is the score of baselines on VMCBench (%).

Backbone Model	**Base Acc.**	**+ ELG (Ours)**	**Improvement (Δ)**	**Reasoning Δ**	**Doc&Chart Δ**
LLaVA-1.5-7B	58.2	**64.0**	**+5.8**	+6.2	+5.5
Qwen2-VL-7B	80.9	**87.1**	**+6.2**	+6.8	+6.0
InternVL-2-8B	72.5	**78.6**	**+6.1**	+6.5	+5.8
Qwen2.5-VL-7B	83.9	**88.6**	**+4.7**	+5.9	+6.4
Qwen3-VL-8B	86.3	**91.6**	**+5.3**	+5.5	+4.8
Average Improvement	–	–	**+5.6**	**+6.2**	**+5.7**

6 Conclusion

This work has formalized Attentional Drift in VLMs, a phenomenon where letterboxing artifacts act as 'attentional black holes' due to a Dual Trap of position bias and feature stagnation. To mitigate this, we have proposed ELG, a training-free framework that rectifies hidden states during inference using spectral and geometric energy priors. Extensive evaluations on VMCBench demonstrate that ELG effectively rehabilitates various VLMs, notably enabling 7B-scale models to surpass 72B-scale and proprietary competitors like GPT-4o. Future work will extend ELG to video-language modeling and robust pre-training. Overall, our approach provides a principled and efficient lens for steering the latent dynamics of VLMs toward reliable multimodal reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for code completion, unit test generation, and language editing. We have not used generative AI tools for formulating research hypotheses or drafting core scientific arguments. All generated code snippets were independently executed, unit-tested, and verified for correctness by the authors. We take full responsibility for the final content and technical artifacts presented in this work.

ETHICS STATEMENT

This work focuses on foundational algorithms and is evaluated entirely on standard, publicly available benchmark datasets. To the best of our knowledge, this work does not involve human participants, proprietary personal data, or direct applications with potential negative societal impacts. All scientific sources and baseline works have been properly cited.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the Energy Landscape Guidance (ELG) framework, including its mathematical formulation, inference-time latent rectification procedure, and experimental configurations. The proposed Energy Scoring Module and Latent Rectification Module are described in Section 3, while Appendix A provides theoretical foundations and derivations for the hidden-state dynamical system, composite energy formulation, and gradient-based rectification procedure. Appendix C provides detailed implementation settings, the VMCBench evaluation protocol, hyperparameter sensitivity analyses, and cross-architecture experiments. We evaluate ELG on VMCBench, comprising 9,450 multiple-choice samples from 20 VQA datasets covering General Scene Understanding, Complex Reasoning, Text-Rich & OCR, and Structured Document Understanding. All experiments follow a consistent evaluation protocol, with efficiency measurements conducted under the same hardware and implementation settings. The complete source code and configuration files will be released upon publication to enable independent verification and future research.

REFERENCES

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4599–4603, 2023. 2

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969, 2021. 1

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 3

Weijue Bu, Guan Yuan, and Guixian Zhang. Conscious gaze: Adaptive attention mechanisms for hallucination mitigation in vision-language models. *arXiv preprint arXiv:2512.05546*, 2025. 3, 8

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 9

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018. 3

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. *arXiv preprint arXiv:2604.03316*, 2026. 3

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019. 3

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 3

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 3

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 3

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 2

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023. 1

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024. 1

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024. 1, 2

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025. 3

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. 3

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 3

Lin Sha, Haiyun Guo, Tao Wang, Cong Zhang, Min Huang, Jinqiao Wang, and Qinghai Miao. St-prune: Training-free spatio-temporal token pruning for vision-language models in autonomous driving. *arXiv preprint arXiv:2604.19145*, 2026. 3

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024. 1

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 1

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023. 3

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. 3

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3, 9

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, et al. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*, 2025a. 3

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431, 2026. 3

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29580–29590, 2025b. 6

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 3

A Theoretical Foundations of Energy Landscape Guidance

In this section, we provide a theoretical interpretation of the Energy Landscape Guidance (ELG) framework introduced in the main text. Our goal is to clarify how the dynamic evolution flux, boundary-aware geometric prior, and semantic consistency constraint jointly define a pathological energy that can be minimized through local hidden-state rectification.

A.1 Hidden-State Evolution as a Discrete Dynamical System

Consider an L-layer Transformer. For visual token t, the hidden state at layer l evolves according to

$$\mathbf{h}_t^l = \mathbf{h}_t^{l-1} + \mathcal{F}^{(l)}\left(\mathbf{h}_t^{l-1}, \mathcal{H}_{/t}^{l-1}\right), \tag{11}$$

where $\mathcal{F}^{(l)}$ denotes the transformation induced by the l-th Transformer block and $\mathcal{H}_{/t}^{l-1}$ denotes the contextual states of the remaining tokens.

This sequence of layer-wise representations defines a discrete trajectory through the pretrained representation space. Informative tokens typically continue to acquire contextual information across successive layers, whereas semantically uninformative or redundant tokens may exhibit substantially weaker changes and approach a nearly stationary trajectory.

Under a continuous-depth interpretation, the layer index can be viewed as a discrete approximation to time,

$$\frac{d\mathbf{h}_t(l)}{dl} = \mathcal{V}\left(\mathbf{h}_t(l), \mathcal{H}_{/t}(l)\right), \tag{12}$$

where $\mathcal{V}$ denotes the effective representation-evolution field. A small layer-wise velocity corresponds to weak representational evolution. This interpretation motivates the evolution-flux statistic used by ELG without requiring the continuous system to be explicitly solved.

A.2 Spectral Interpretation of Evolution Flux

Following Sec. 3.1.1 of the main text, consider the most recent K hidden states of token t,

$$\mathcal{T}_t = \left[\mathbf{h}_t^{l-K+1}, \ldots, \mathbf{h}_t^l\right]. \tag{13}$$

Its discrete Fourier transform is

$$\mathbf{X}_t[\omega] = \sum_{k=0}^{K-1} \mathbf{h}_t^{l-k} \exp\left(-j\frac{2\pi}{K}\omega k\right). \tag{14}$$

The zero-frequency component summarizes the stationary component of the trajectory, whereas the non-zero-frequency components characterize variation across layers. We therefore define the evolution flux as

$$\Phi_{\text{flux}}(\mathbf{h}_t^l) = \log\left(1 + \frac{1}{K}\sum_{\omega=1}^{K-1} \|\mathbf{X}_t[\omega]\|_2^2\right). \tag{15}$$

A large Φ_{flux} indicates persistent layer-wise refinement, whereas a small value indicates a nearly stationary trajectory. By Parseval's identity, the energy contained in non-zero frequencies measures deviation from the layer-wise mean. Consequently, if

$$\mathbf{h}_t^{l-k} \approx \bar{\mathbf{h}}_t \qquad \text{for all } k, \tag{16}$$

then the non-zero-frequency coefficients approach zero and $\Phi_{\text{flux}}(\mathbf{h}_t^l)$ becomes small. Thus, evolution flux provides a compact diagnostic of representational stagnation over a multi-layer window rather than relying on a single adjacent-layer difference.

Importantly, Φ_{flux} is a refinement score: larger values correspond to stronger representation evolution. To obtain an objective whose minimization is consistent with reducing stagnation, ELG converts this score into the stagnation energy

$$\mathcal{E}_{\text{stag}}(\mathbf{h}_t^l) = -\log\left(\Phi_{\text{flux}}(\mathbf{h}_t^l) + \epsilon\right), \tag{17}$$

where $\epsilon > 0$ ensures numerical stability. Since $\mathcal{E}_{\text{stag}}$ decreases monotonically with Φ_{flux}, low-flux trajectories receive a larger pathological energy, whereas actively evolving representations incur a smaller penalty.

A.3 Boundary-Aware Geometric Prior

Low evolution flux alone does not imply that a token is necessarily harmful. The central hypothesis of ELG is that stagnation becomes particularly problematic when it coincides with a vulnerable sequential position. Accordingly, the geometric prior should reflect both latent stagnation and token position.

Rather than using a discrete stagnant-set membership inside the optimization objective, we define a differentiable stagnation gate

$$s_t^l = \sigma\left(\frac{\tau - \Phi_{\text{flux}}(\mathbf{h}_t^l)}{\delta}\right), \tag{18}$$

where τ is an adaptive threshold estimated from the current flux distribution, $\delta > 0$ controls the transition smoothness, and $\sigma(\cdot)$ is the sigmoid function. When $\Phi_{\text{flux}}(\mathbf{h}_t^l) \ll \tau$, the gate approaches one; for actively evolving tokens, it approaches zero.

The Boundary-aware Geometric Prior is then defined as

$$\Psi_{\text{geo}}(\mathbf{h}_t^l, t) = s_t^l \ln\left(1 + \text{pos}(t)\right), \tag{19}$$

where $\text{pos}(t)$ denotes the relative sequential position of token t. This formulation has two complementary effects. The stagnation gate assigns larger weight to tokens exhibiting weak layer-wise evolution, while the logarithmic positional factor increases the penalty when such tokens occur later in the visual sequence. The logarithm prevents excessive positional amplification.

Unlike a hard membership rule, s_t^l remains differentiable with respect to the hidden state through Φ_{flux}. Therefore, Ψ_{geo} not only identifies positionally vulnerable stagnant tokens but also contributes directly to the gradient used for latent rectification.

A.4 Composite Energy and Latent Rectification

The Energy Scoring Module combines the three terms defined in the main text:

$$\mathcal{E}(\mathbf{h}_t^l) = \alpha\mathcal{E}_{\text{stag}}(\mathbf{h}_t^l) + \beta\Psi_{\text{geo}}(\mathbf{h}_t^l, t) + \gamma\Omega_{\text{cons}}(\mathbf{h}_t^l). \tag{20}$$

Here, $\mathcal{E}_{\text{stag}}$ penalizes insufficient representation evolution, Ψ_{geo} increases this pressure for stagnant tokens at vulnerable sequential positions, and Ω_{cons} constrains the rectified state relative to its original representation. The term "energy" therefore refers to a pathological score that ELG explicitly seeks to decrease rather than to a separate probabilistic energy-based model.

For layer l, the vanilla hidden state is first computed as

$$\mathbf{h}_t^l = f^{(l)}\left(\mathcal{H}^{(l-1)*}\right)_t. \tag{21}$$

The Latent Rectification Module then performs a single local intervention:

$$\mathbf{h}_t^{l*} = \mathbf{h}_t^l - \eta\nabla_{\mathbf{h}_t^l}\mathcal{E}(\mathbf{h}_t^l). \tag{22}$$

For sufficiently small η, a first-order Taylor expansion gives

$$\mathcal{E}\left(\mathbf{h}_t^l - \eta\nabla_{\mathbf{h}_t^l}\mathcal{E}\right) = \mathcal{E}(\mathbf{h}_t^l) - \eta\left\|\nabla_{\mathbf{h}_t^l}\mathcal{E}(\mathbf{h}_t^l)\right\|_2^2 + O(\eta^2). \tag{23}$$

Hence, whenever the local gradient is non-zero and η is sufficiently small, the intervention locally decreases the pathological energy.

The direction of this update is consistent with the intended rectification. Because

$$\frac{\partial\mathcal{E}_{\text{stag}}}{\partial\Phi_{\text{flux}}} = -\frac{1}{\Phi_{\text{flux}} + \epsilon} < 0, \tag{24}$$

reducing the stagnation energy favors increasing the evolution flux rather than suppressing it. Meanwhile, the differentiable gate in Ψ_{geo} strengthens this correction when low-flux tokens occur at vulnerable sequential positions. The semantic term counterbalances both effects by discouraging excessive deviation from the original representation.

A.5 Semantic Consistency as a Local Stability Constraint

The rectification step should modify pathological hidden-state behavior without arbitrarily rotating the pretrained representation. ELG therefore uses the cosine-distance regularizer from the main text:

$$\Omega_{\text{cons}}(\mathbf{h}) = 1 - \frac{\mathbf{h}^\top \hat{\mathbf{h}}}{\|\mathbf{h}\|_2 \|\hat{\mathbf{h}}\|_2}, \tag{25}$$

where $\hat{\mathbf{h}}$ denotes the corresponding hidden state before rectification.

The cosine term penalizes changes in representation direction. In particular,

$$\Omega_{\text{cons}}(\mathbf{h}) \geq 0, \tag{26}$$

with equality when $\mathbf{h}$ and $\hat{\mathbf{h}}$ are positively collinear. Consequently, this term acts as a directional anchor during the local intervention: ELG can alter the hidden-state dynamics while discouraging large angular deviations from the pretrained representation.

The three components therefore have distinct roles. $\mathcal{E}_{\text{stag}}$ determines whether the latent trajectory exhibits insufficient evolution, Ψ_{geo} modulates the rectification according to positional vulnerability, and Ω_{cons} limits semantic distortion introduced by the intervention.

A.6 Interpretation of the Dual Trap

The preceding formulation provides a unified interpretation of the two factors described as the *Dual Trap* in the main text.

First, a low value of Φ_{flux} indicates that a token has undergone little representational change over the recent layer window. This produces a large $\mathcal{E}_{\text{stag}}$, identifying a dynamically stagnant representation. Second, the geometric prior increases when such stagnation coincides with a later sequential position, where the positional effect discussed in the main text can make its influence more consequential. Neither factor alone is assumed to be sufficient; the strongest rectification pressure arises from their conjunction.

ELG therefore combines (i) a dynamic diagnostic of representational stagnation, (ii) a differentiable position-dependent modulation of that stagnation, and (iii) a semantic constraint on the resulting hidden-state intervention. This decomposition captures the interaction between feature stagnation and positional vulnerability without requiring an externally specified padding mask.

A.7 Summary

The theoretical interpretation above follows the same objective used by the main method. At each rectified layer, ELG:

1. computes the vanilla hidden states;
2. estimates Φ_{flux} from the recent hidden-state trajectory;
3. converts the evolution score into $\mathcal{E}_{\text{stag}}$;
4. evaluates the differentiable stagnation gate s_t^l and the boundary-aware prior Ψ_{geo};
5. evaluates the semantic consistency term Ω_{cons}; and
6. applies the single-step latent rectification

$$\mathbf{h}_t^{l*} = \mathbf{h}_t^l - \eta \nabla_{\mathbf{h}_t^l} \mathcal{E}(\mathbf{h}_t^l). \tag{27}$$

The next Transformer layer then operates on the rectified hidden states $\mathcal{H}^{l*}$. ELG therefore retains the pretrained model parameters while performing a local inference-time intervention that decreases pathological stagnation energy, strengthens correction at vulnerable positions, and constrains semantic drift.

B Additional Experimental Results

B.1 Implementation Details

All experiments were implemented in PyTorch 2.1, with Qwen2.5-VL used as the primary backbone. Experiments were conducted on a computing cluster equipped with eight NVIDIA H20 GPUs. For consistency, all efficiency measurements, including the reported throughput of 62 images/sec, were obtained on the same H20-based hardware under identical evaluation settings.

B.2 Hyperparameter Sensitivity

Unless otherwise specified, we use a single configuration, $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 0.2$, and $\eta = 0.02$, which is selected on Qwen2.5-VL-7B and subsequently kept fixed across experiments. Rather than performing model-specific hyperparameter tuning, we conduct one-dimensional perturbations around this configuration to examine the robustness of ELG. All experiments in this section are evaluated on VMCBench using Qwen2.5-VL-7B.

Sensitivity to the Rectification Step Size (η). We first vary the rectification step size while keeping the energy weights fixed at $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.2$.

As shown in Table 6, ELG improves over the unrectified baseline for a broad range of step sizes. Performance increases from 83.9% at $\eta = 0$ to 88.6% at $\eta = 0.02$, while nearby values such as $\eta = 0.01$ and $\eta = 0.05$ achieve 87.2% and 87.8%, respectively. The improvement therefore does not depend on a narrowly tuned step size. Only an excessively large update ($\eta = 0.2$) degrades performance, suggesting that overly aggressive rectification can perturb the hidden representations too strongly. Based on these results, we use $\eta = 0.02$ throughout the remaining experiments.

Table 6: Sensitivity to the rectification step size η on VMCBench. The energy weights are fixed to $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.2$.

Step Size η	0.0 (Base)	0.005	0.01	**0.02**	0.05	0.1	0.2
Avg. Accuracy (%)	83.9	85.1	87.2	**88.6**	87.8	85.5	81.5
Δ Performance	–	+1.2	+3.3	**+4.7**	+3.9	+1.6	-2.4

Sensitivity to the Energy Weights (α, β, γ). We next vary each energy weight independently while keeping the remaining parameters fixed. The results are summarized in Tables 7-9.

Table 7: Sensitivity to the evolution-flux weight α.

Weight Value (α)	0.1	0.5	**1.0**	2.0	5.0
Accuracy (%) ($\beta = 0.5, \gamma = 0.2$)	83.5	86.8	**88.6**	87.9	85.2

Table 8: Sensitivity to the geometric-prior weight β.

Weight Value (β)	0.1	**0.5**	1.0	2.0	5.0
Accuracy (%) ($\alpha = 1.0, \gamma = 0.2$)	85.8	**88.6**	88.2	87.0	84.1

Table 9: Sensitivity to the semantic-consistency weight γ.

Weight Value (γ)	0.1	0.2	**0.5**	1.0	2.0	5.0
Accuracy (%) ($\alpha = 1.0, \beta = 0.5$)	86.5	**88.6**	**88.6**	88.0	86.3	82.5

Overall, the three terms exhibit relatively broad effective ranges rather than isolated optima. For α, performance remains high around 1.0–2.0, whereas very small or large weights are less effective. For β, both 0.5 and 1.0 produce similar results, with only a 0.4-point difference. Similarly, $\gamma = 0.2$ and

$\gamma = 0.5$ both achieve 88.6%, and $\gamma = 1.0$ remains close at 88.0%. These results indicate that ELG is tolerant to moderate variations in the relative weighting of its energy components and does not require precise tuning of individual coefficients.

Cross-Architecture Hyperparameter Transferability. To further assess whether ELG requires architecture-specific tuning, we determine the parameter set $(\alpha, \beta, \gamma, \eta) = (1.0, 0.5, 0.2, 0.02)$ using only Qwen2.5-VL-7B and directly apply the same configuration to other VLM architectures without additional tuning.

B.3 Performance in Challenging Scenarios

Small Object Scenarios The perception of spatially constrained targets (small objects) and intricate environments remains a significant challenge for current VLMs, as these scenarios often lack strong visual cues, leading to 'attentional traps'. We qualitatively analyze these behaviors in Figure 5.

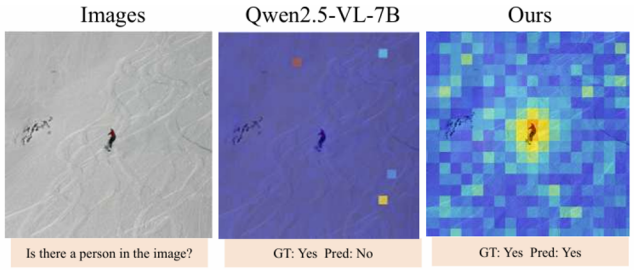


Figure 5: Qualitative analysis in complex and small-object scenarios. We compare the attention maps of the vanilla Qwen2.5-VL-7B (baseline) and our ELG. Left (Small Object): In a vast, low-contrast background, the baseline's attention is diffused, failing to 'see' the skier. ELG successfully captures the diminutive target.

As illustrated in the small-object scenario (left), the vanilla Qwen2.5-VL-7B exhibits 'attention diffusion', where the model's focus is swallowed by the vast, low-contrast snowy background, resulting in a failure to identify the skier (predicting 'No'). By applying ELG, the spectral flux analysis (Φ_{flux}) effectively suppresses the uninformative background tokens and amplifies the signal of the diminutive target, successfully rectifying the prediction.

Degraded Scenes To further evaluate the robustness of ELG against common low-level visual artifacts, we examine its performance in 'degraded' scenarios involving the addition of additive white noise across the entire image. As illustrated in Figure 6, the introduction of stochastic perturbations triggers a significant degradation in the vanilla Qwen2.5-VL-7B. In the original image, the baseline model correctly identifies the scene. However, when the input is corrupted by white noise, the attention maps reveal a significant 'attentional drift,' where the focus is scattered across the noisy background instead of being concentrated on salient objects. This disruption acts as a 'black hole' that drains representational energy from the core features, causing the model to lose fine-grained details (e.g., the frisbee) and hallucinate a generic relationship (predicting 'a man is chasing another man').

In contrast, our ELG framework demonstrates remarkable stability against such pixel-level noise. By leveraging the Energy Scoring Module, ELG effectively filters out the high-frequency energy fluctuations caused by the white noise (Φ_{flux}) and maintains the structural integrity of the attention map. As a result, the rectified attention remain tightly focused on the salient athletes and the frisbee, ensuring that ELG consistently yields the correct description across both clean and noisy inputs.

C Limitations and Future Work

While Energy Landscape Guidance effectively rehabilitates VLMs from attentional drift in static images, its current formulation has two coupled limitations that define our future research directions. First, our evaluation is primarily focused on static benchmarks and single-turn VQA. In the context

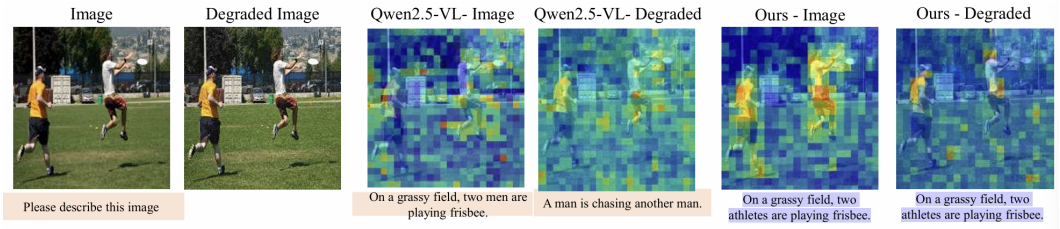


Figure 6: Robustness of ELG under image degradation with additive white noise. Compared with the vanilla Qwen2.5-VL-7B, whose attention becomes more dispersed under the degraded input and leads to an incorrect description ("A man is chasing another man"), ELG maintains greater attention on the salient athletes and the frisbee. Consequently, ELG preserves the correct description ("two athletes are playing frisbee") under both the original and degraded inputs.

of long-video reasoning, the model encounters a more complex, time-varying form of "Feature Stagnation" where repetitive frames or temporal redundancies act as dynamic attractors. Although the universal hyperparameter set (α, β, γ) demonstrates robust generalization across diverse image-based architectures, these static coefficients may not optimally capture the fluctuating informational flux of long-duration temporal sequences.

Consequently, a second limitation is the heuristic nature of the current energy calibration. While ELG achieves "plug-and-play" deployment without per-model tuning, it still lacks a fully automated mechanism to dynamically adjust rectification intensity in response to real-time entropy changes in video streams. Future work will focus on extending ELG to the temporal dimension and developing an adaptive energy controller that autonomously calibrates the guidance strength based on the instantaneous manifold geometry, thereby providing a more resilient solution for complex, multi-modal temporal reasoning.