Paying More Attention to Visual Tokens in Self-Evolving Large Multimodal Models

Shravan Venkatraman[1], Ritesh Thawkar[1], Omkar Thawakar[1], Rao Muhammad Anwer[1,2], Hisham Cholakkal[1], Salman Khan[1,3], and Fahad Shahbaz Khan[1,4]

[1] Mohamed bin Zayed University of Artificial Intelligence [2] Aalto University
[3] Australian National University [4] Linköping University

Abstract. Recently, self-evolving large multimodal models (LMMs) have received attention for improving visual reasoning in a purely unsupervised setting. However, multi-role self-play and self-consistency reward schemes in existing self-evolving LMMs optimize answer agreement without ensuring the decoder attends to visual content, relying instead on statistical language priors to produce self-consistent outputs. This leads to a persistent failure mode we term visual under-conditioning, where the decoder relies on language priors rather than the image during generation, manifesting as insufficient attention to visual tokens. As a result, current self-evolving LMMs struggle on vision–language understanding tasks such as image captioning and visual question answering. To address this, we propose VISE (Visual Invariance Self-Evolution), a purely unsupervised self-evolving framework that directly regularizes the model's visual conditioning policy through two complementary invariance-based rewards: a geometric invariance reward that enforces spatial consistency under known transformations, and a semantic invariance reward that penalizes evidence-agnostic generation by requiring the model to recognize the absence of evidence when predicted regions are perturbed. VISE operates within a single model without specialist roles, external reward models, or annotations, and is trained on raw unlabeled images. Experiments on 18 benchmarks demonstrate the efficacy of our approach. Using Qwen3-VL-2B as the base model, VISE achieves gains of +16.85 CIDEr on COCO and +19.66 CIDEr on TextCaps, reduces object hallucination by 5.0 Chair-I points, and generalizes across four model families and scales. Our code and models are available at https://mbzuai-oryx.github.io/VISE/.

1 Introduction

Recent work on self-evolving large multimodal models (LMMs) has demonstrated that models can improve their visual reasoning capabilities in a purely unsupervised manner, without relying on human-annotated supervision or externally verified reward models [6,28–30]. These approaches frame multimodal reasoning as a self-improving process, instantiating proposer–solver or questioner–reasoner roles within a shared backbone and reinforcing multi-sample agreement, trajectory-aware feedback, and tool-assisted verification to promote more reliable reasoning [6, 14, 28, 29]. Despite their promise, however, these methods optimize for

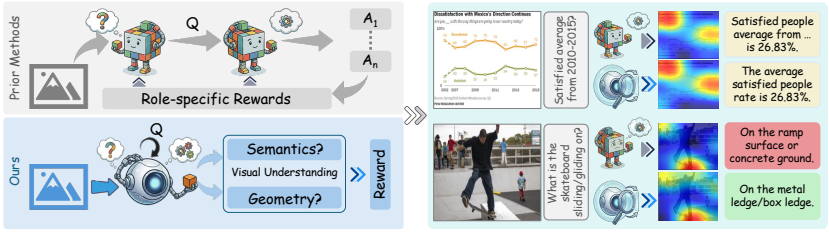

Fig. 1: Overview of VISE versus prior self-evolving methods. [6, 28–30] *Left:* Prior methods use separate roles and answer-consistency objectives, while VISE uses a single model with geometric and semantic invariance rewards. *Right:* On chart queries with limited visual dependence, both methods answer correctly; on real-scene understanding, prior methods default to plausible language priors ('ramp surface") rather than visual evidence ('metal ledge"), whereas VISE identifies the correct region after invariance-based training. Additional examples are in the suppl. material.

answer agreement without ensuring the decoder attends to the actual visual content, relying instead on statistical language priors to produce self-consistent outputs. This leads to a persistent failure mode we term *visual under-conditioning*, which causes hallucination, modality bypass, and unstable visual interpretation even when the vision encoder produces accurate representations. Concretely, this manifests as insufficient attention to visual tokens during decoding, where generations are driven more by language priors than by the image being described.

This limitation arises from two structural features common to existing self-evolving frameworks. First, the Proposer–Solver setup forms an implicit minimax game: the roles are optimized jointly with opposing objectives, making training unstable over long horizons. In practice, one role often dominates; Proposer collapses to trivial or degenerate queries that guarantee agreement, or the Solver overfits to the Proposer's distribution and fails to generalize, driving the system into local minima that are difficult to correct without external intervention. Second, framing the reward around answer correctness implicitly assumes that self-consistency implies correct visual grounding. A decoder exploiting language priors can achieve high answer agreement through statistical co-occurrences alone, leaving visual under-conditioning unaddressed or even reinforced.

As illustrated in Fig. 1, this failure mode is directly observable. On chart-based or structured scientific queries, prior self-evolving methods produce the same correct answers as ours, suggesting that the required visual cues may be sparse or quickly resolved, after which the decoder can proceed largely from its internal representations. However, on natural scene understanding queries such as identifying what a skateboard is actually resting on, these methods default to statistically common associations (ramp surface, concrete ground), indicating that visual evidence is not being robustly integrated throughout the generation process. In contrast, correctly resolving such cases requires the decoder to remain conditioned on the specific image content, exposing the gap between answer consistency and genuine visual interpretation. Fig. 2 corroborates this pattern at

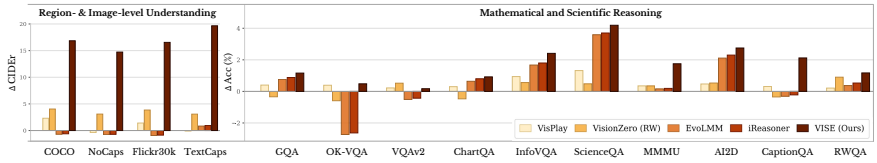

Fig. 2: Performance delta relative to the base model across captioning and reasoning benchmarks (Qwen3-VL-2B). Delta CIDEr (left) and delta accuracy (right) for five self-evolving methods. Methods trained on math- and science-centric images improve structured reasoning (e.g., EvoLMM +3.59 on ScienceQA) but regress on captioning, whereas VISE achieves the largest captioning gains (+16.85 COCO, +19.66 TextCaps) while remaining competitive on reasoning tasks.

scale. Prior self-evolving methods trained on science- and math-centric images show gains on reasoning benchmarks, yet perform at or below the base model on captioning and region-level description tasks that require detailed visual conditioning. This disparity validates the hypothesis that improvements in answer agreement do not translate into stronger visual interpretation, and that visual under-conditioning persists in existing self-evolving frameworks.

To directly address this, we propose **VISE** (**V**isual **I**nvariance **S**elf-**E**volution), a purely unsupervised self-evolving framework that regularizes the model's visual conditioning policy rather than answer agreement. Unlike prior multi-role formulations, VISE operates within a single model, motivated by the observation that a well-pretrained LMM already possesses sufficient knowledge to formulate meaningful queries about its visual content. Its training signal comes from two complementary invariance-based rewards: a *geometric invariance* reward that enforces spatial consistency under known transformations, and a *semantic invariance* reward that requires the model to recognize the absence of evidence when predicted regions are perturbed. Training on raw images with no annotations, metadata, or external reward models, VISE achieves gains of up to +16.85 CIDEr on COCO and +19.66 CIDEr on TextCaps, reduces object hallucination (Chair-I) by 5.0 points, and improves consistently across VQA and reasoning benchmarks, demonstrating that stronger visual conditioning generalizes across tasks rather than trading off against them. Mechanistically, VISE achieves this by increasing attention to visual tokens across decoder layers during generation, reflecting the shift from language-prior-driven to image-conditioned decoding.

In summary, our main contributions are as follows:

- We introduce VISE, a single-model, fully unsupervised self-evolving framework that directly addresses visual under-conditioning in LMMs by regularizing the model's visual conditioning policy rather than its answer agreement, requiring no annotations, metadata, or external reward models.
- We propose two complementary invariance-based reward signals: geometric invariance and semantic invariance via regional perturbation (ghosting), that jointly enforce spatial consistency and evidence sensitivity, providing a purely self-supervised objective defined entirely from the model's own predictions.

– We empirically validate VISE across 18 benchmarks spanning image captioning, VQA, reasoning, and hallucination on multiple model scales and four backbone families, demonstrating consistent and substantial improvements over all self-evolving baselines with no tradeoffs between task groups.

2 Related Work

Early work on self-improving LMMs aimed to reduce reliance on human-annotated data through internal preference and alignment signals. Tan *et al.* [27] introduce image-driven self-questioning with diffusion-based rejection for DPO optimization. Wang *et al.* [31] propose an in-context self-critic with visual metrics to construct preference pairs without external models. Liu *et al.* [15] study multimodal self-evolution from a reinforcement learning perspective and introduce entropy-driven exploration to mitigate saturation. While these approaches reduce annotation dependence, they still rely on structured supervision or curated preference construction, limiting fully autonomous self-improvement.

More recent work has shifted to fully unsupervised self-play frameworks with internally generated rewards, yielding strong gains on structured reasoning benchmarks. EvoLMM [29] introduces a Proposer–Solver formulation optimized with continuous self-consistency rewards to enable co-evolution of question generation and reasoning without annotations. Subsequent extensions refine this paradigm: iReasoner [28] incorporates trajectory-aware rewards, VisPlay [6] promotes diversity and difficulty to prevent collapse, and Agent0-VL [14] integrates tool-grounded self-verification. C2-Evo [3] and DoGe [11] further address training instabilities through co-evolutionary data loops and role decoupling mechanisms.
Limitations. Despite their methodological diversity and strong performance on structured reasoning benchmarks, these approaches optimize answer correctness or reasoning consistency as the primary objective. This implicitly assumes that self-consistent outputs reflect improved visual understanding, which is an assumption that breaks down under visual under-conditioning, where a model can remain self-consistent and even correct while relying on statistical language priors rather than visual evidence. VISE addresses this directly by replacing answer-agreement rewards with invariance-based rewards that regularize the model's visual-conditioning policy itself, operating within a single model on raw unlabeled images without specialist roles or external reward models.

3 Method

Problem Formulation. Let $\mathcal{X} = \{x\}$ denote an unlabeled collection of images, with no accompanying queries, bounding box annotations, or category labels. At each training step, the model π_θ operates in a self-questioning regime: it first generates a natural-language localization query q by interrogating image x, then predicts a bounding box $B = (x_1, y_1, x_2, y_2)$ locating the queried object under the same query. Both query generation and bounding box prediction are performed by the same single policy, without separate specialist roles, as a

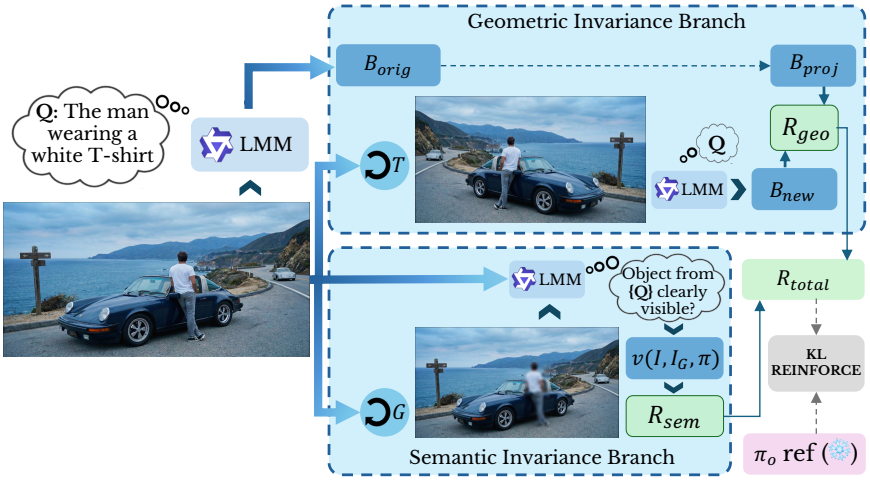

Fig. 3: Overview of the VISE self-evolving framework. Given a raw unlabeled image, the model first generates a localization query and predicts a bounding box B_{orig}. The Geometric Invariance Branch applies a spatial transformation $\mathcal{T}$, predicts B_{new} on the transformed view, and computes $\mathcal{R}_{\text{geo}}$ as the GIoU between B_{new} and the projected box B_{proj} to enforce spatial consistency across views. The Semantic Invariance Branch ghosts the predicted region via blurring and assigns $\mathcal{R}_{\text{sem}}$ only if the model detects the object before perturbation and not afterward, penalizing evidence-agnostic generation. The combined reward is optimized with KL-regularized REINFORCE against a frozen reference policy π_o, without annotations, external reward models, or specialist roles.

well-pretrained LMM already possesses sufficient visual knowledge to formulate meaningful queries for its own content. All spatial coordinates are represented in a normalized space $[0, S]^4$, where $S = 1000$, such that a pixel-space coordinate c_{pix} along dimension D maps to $\tilde{c} = (c_{\text{pix}}/D) \cdot S$.

We use the Qwen3-VL family [2] as the base backbone, freezing the vision encoder and updating the multimodal projector, feed-forward layers, and decoder attention projections. This is motivated by the nature of the failure mode: the vision encoder already produces strong visual representations, and the problem lies in how the decoder projects and utilizes them. Propagating noisy, unsupervised reward gradients into the encoder would risk destabilizing representations that are already high quality, without addressing the locus of failure.

Geometric Invariance Reward. Visual under-conditioning manifests most directly in localization instability: a decoder that disengages with visual image evidence will produce predictions that are inconsistent with the actual spatial structure of the scene, and in particular will fail to maintain coherent localization when the image undergoes a known geometric transformation. We exploit this as a self-supervised training signal. If the model correctly conditions its localization on what it sees, then its predicted box on a geometrically transformed image should correspond precisely to the analytically projected version of its prediction

on the original. Deviation from this consistency is evidence of visual under-conditioning, and we penalize it directly.

At each training step, the model generates a query q by conditioning on image x, producing a short natural-language description of a prominent and spatially unambiguous object in the scene. The model then predicts a bounding box B_{orig} on the image under query q. A geometric transformation $\mathcal{T}$ is sampled uniformly from three types: affine (rotation $\theta \sim \mathcal{U}(-10°, 10°)$, scale $s \sim \mathcal{U}(0.9, 1.1)$, translation $(\delta_x, \delta_y) \sim \mathcal{U}(-50, 50)^2$), crop (ratio $\rho \sim \mathcal{U}(0.8, 1.0)$, resized to original resolution), or horizontal flip. Each is described by a known 3×3 homogeneous matrix M. The model then predicts a second box B_{new} on the transformed image $x' = \mathcal{T}(x)$ under the same query q. The expected box under the transformation is computed by lifting the four corners of B_{orig} to homogeneous coordinates and applying M as $\mathbf{c}'_i = M\mathbf{c}_i$; the axis-aligned box of the resulting corners gives the projected box B_{proj}. The geometric invariance reward is:

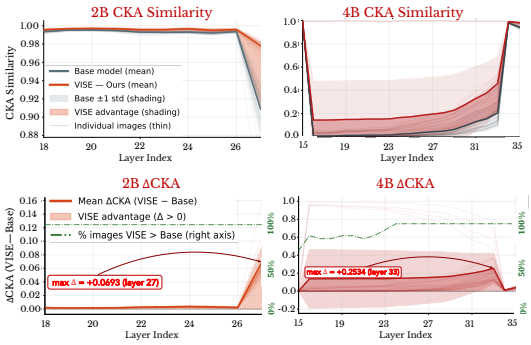

Fig. 4: Per-layer CKA similarity between original and geometrically augmented views in Qwen3-VL decoder layers. On 2B, VISE gains are confined to final layers (peak $\Delta = +0.069$ at layer 27) with 100% win-rate. On 4B, gains span layers 19–33 (peak $\Delta = +0.253$), with win-rate increasing from ∼60% at layer 15 to 100% beyond layer 25.

$$\mathcal{R}_{\text{geo}} = \frac{\text{GIoU}(B_{\text{proj}},\, B_{\text{new}}) + 1}{2} \tag{1}$$

where GIoU is the Generalized Intersection over Union [23]:

$$\text{GIoU}(B_1, B_2) = \text{IoU}(B_1, B_2) - \frac{|\mathcal{C}| - |B_1 \cup B_2|}{|\mathcal{C}|} \tag{2}$$

with $\mathcal{C}$ denoting the smallest axis-aligned box enclosing both B_1 and B_2. The linear normalization in Eq. (1) maps GIoU $\in [-1, 1]$ to $\mathcal{R}_{\text{geo}} \in [0, 1]$. This reward is maximized when the model's localization on the transformed view agrees precisely with the geometric projection of its original prediction, and degrades smoothly as spatial consistency deteriorates. In this way, $\mathcal{R}_{\text{geo}}$ directly targets the spatial dimension of visual under-conditioning: such a model cannot maintain such consistency across views and is therefore penalized, even if its individual predictions appear plausible in isolation.

Figure 4 provides representational evidence that $\mathcal{R}_{\text{geo}}$ achieves its intended effect. We compute per-layer Centered Kernel Alignment (CKA) similarity [10]

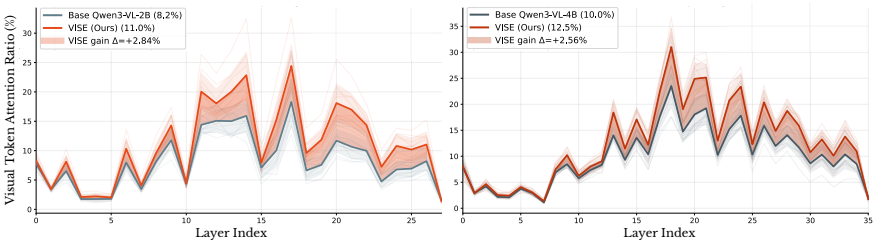

Fig. 5: Generation-time visual attention per transformer layer for Base and VISE models on Qwen3-VL-2B (left) and Qwen3-VL-4B (right). VISE-trained models (orange) consistently assign more attention to image tokens across mid-to-late decoder layers where semantic generation occurs, with mean gains of +2.84% and +2.56% respectively and per-sample peaks of up to +5.09% in layers 15–25. The effect is consistent across both model scales, aligning with our claim that the semantic invariance reward strengthens visual conditioning during generation.

between representations of original and geometrically augmented views for the base model and VISE across 100 random COCO images. On Qwen3-VL-2B, gains are confined to the final decoder layers, localizing geometric under-conditioning to the stages where generation decisions are formed. On Qwen3-VL-4B, the ΔCKA advantage is distributed more broadly across the decoder, consistent with the scale effects in Table 1: concentrated failures yield more direct downstream gains under invariance-based correction.

Semantic Invariance Reward. Geometric consistency is a necessary but insufficient condition for faithful visual conditioning. A model could achieve high $\mathcal{R}_{\text{geo}}$ by predicting large, spatially stable regions without its predictions being meaningfully driven by the semantic content they enclose. Hence, we also address the complementary dimension of visual under-conditioning: *evidence sensitivity*. A model whose responses are conditioned on the image should recognize that removing the predicted region removes the evidence for the queried object. If instead the decoder is shortcutting from language priors, it would remain insensitive to that removal. We reward the opposite: the model should judge the object as visible when the region is intact, and as absent when it is obscured.

We do this by introducing a regional perturbation procedure termed *ghosting*. Given the predicted box B_{orig}, the corresponding pixel region in the original image x is identified and its contents replaced by a Gaussian-blurred version with kernel $\sigma = 25.0$, producing a ghosted image $\tilde{x}$ in which the localized region is visually degraded while the surrounding context is fully preserved. The model then assesses the visibility of the queried object under q on both x and $\tilde{x}$ via greedy decoding, yielding binary judgments $v = \texttt{vis}(x, q) \in \{0, 1\}$ and $\tilde{v} = \texttt{vis}(\tilde{x}, q) \in \{0, 1\}$. The semantic invariance reward is:

$$\mathcal{R}_{\text{sem}} = \begin{cases} 1.0 & \text{if } v = 1 \text{ and } \tilde{v} = 0 \\ 0.0 & \text{otherwise} \end{cases} \tag{3}$$

A prediction that is geometrically consistent but semantically arbitrary (enclosing a region that does not contain the queried object) receives $\mathcal{R}_{\text{sem}} = 0$ even

if it satisfies $\mathcal{R}_{\text{geo}}$, ensuring the two signals are complementary and jointly necessary for the model to improve its evidence binding. Figure 5 provides direct evidence of the intended behavioral change. We measure generation-time visual attention, which is the fraction of attention each generated token assigns to image tokens across transformer layers, for the base model and VISE averaged over 100 random COCO images. Unlike prefill attention, which is dominated by the visual embedding layer, generation-time attention captures how the decoder references visual evidence while producing each token. VISE consistently assigns more attention to visual tokens across mid-to-late decoder layers on both model scales, indicating that the semantic invariance reward encourages the decoder to maintain visual conditioning throughout generation rather than reverting to language priors after initial image encoding.

Composite Reward and Optimization. The total reward combines both invariance signals as $\mathcal{R}_t = \lambda_{\text{geo}}\mathcal{R}_{\text{geo}} + \lambda_{\text{sem}}\mathcal{R}_{\text{sem}}$, where $\lambda_{\text{geo}} = \lambda_{\text{sem}} = 0.5$. At each step, the model produces a completion y containing the predicted box coordinates for (x, q). We maintain an exponential moving average baseline $b_t \leftarrow 0.9\, b_{t-1} + 0.1\, \mathcal{R}_t$ to reduce gradient variance, giving advantage $A_t = \mathcal{R}_t - b_t$. Letting $\Delta_t = \log p_\theta(y \mid x, q) - \log p_{\text{ref}}(y \mid x, q)$ denote a KL-like divergence proxy between the current policy and the frozen reference π_{ref}, we optimize:

$$\mathcal{L}(\theta) \;=\; -\,A_t \cdot \log p_\theta(y \mid x, q) \;+\; \beta_t \cdot \Delta_t \tag{4}$$

where the first term is a REINFORCE-style update and the second regularizes against the reference model. We adapt the KL Coefficient β_t dynamically to maintain a target divergence level:

$$\beta_{t+1} = \begin{cases} \beta_t\,(1+\eta) & \text{if } |\Delta_t| > \tau \\ \beta_t\,(1-\eta) & \text{otherwise} \end{cases} \tag{5}$$

where $\tau = 0.020$ is the target divergence budget, $\eta = 0.10$ is the adaptation rate, and β_t is clipped below at 10^{-6}. This tightens regularization when the policy drifts beyond the target and relaxes it when updates are conservative, providing stable self-evolution without a fixed regularization strength.

4 Experiments

Implementation Details. We fine-tune each base model using LoRA [7] while keeping the vision encoder frozen. For the smaller models (2B and 4B), we use rank r=16 and α=32, whereas for the larger models (8B and 32B), we increase the rank to r=32 and α to 64 (dropout 0.05 in all cases). Training uses the AdamW optimizer [17] with weight decay 0.01 and gradient clipping at 1.0. We set the learning rate to 10^{-6} and the KL regularization target to 0.020 (adaptive rate 0.10) for smaller models, and reduce them to 1.5×10^{-7} and 0.004 (adaptive rate 0.15), respectively, for larger models. Equal reward weights of 0.5 are applied to both the geometric and semantic invariance terms. All models are trained for 4000 steps on 8× AMD MI250X GPUs using bfloat16 precision. No question–answer pairs, annotations, metadata, or external reward models are used.

Table 1: Evaluation results on four image captioning benchmarks. C = CIDEr, M = METEOR, R = ROUGE-L. Consistency-based methods trained on math and scientific images (EvoLMM, iReasoner) regress on captioning across all scales; EvoLMM drops −0.70 CIDEr on COCO and −0.94 on Flickr30k at 2B, suggesting that prior-driven generation reinforces language priors rather than correcting them. Conversely, VISE improves CIDEr from 21.54 → 38.39 on COCO and 22.20 → 41.86 on TextCaps at 2B, with no regressions across any dataset or scale.

Method	COCO			NoCaps			Flickr30k			TextCaps		
	C	M	R	C	M	R	C	M	R	C	M	R
Qwen3-VL-2B-Instruct												
Base	21.54	26.31	42.35	19.52	27.36	45.08	26.09	25.70	43.75	22.20	25.31	38.66
VisPlay [6]	23.85 +2.31	26.61 +0.30	42.65 +0.30	19.14 -0.38	27.83 +0.47	45.43 +0.35	27.50 +1.41	25.53 -0.17	44.20 +0.45	22.11 -0.09	25.46 +0.15	39.13 +0.47
VisionZero (CLEVR) [30]	22.67 +1.13	26.34 +0.03	42.88 +0.53	20.16 +0.64	28.01 +0.65	44.23 -0.85	27.78 +1.69	24.89 -0.81	44.65 +0.90	22.82 +0.62	24.49 -0.82	39.31 +0.65
VisionZero (Chart) [30]	21.47 -0.07	26.96 +0.65	43.10 +0.75	20.19 +0.67	26.74 -0.62	44.33 -0.75	26.98 +0.89	26.52 +0.82	42.98 -0.77	21.41 -0.79	25.96 +0.65	39.47 +0.81
VisionZero-RW [30]	25.58 +4.04	27.20 +0.89	43.46 +1.11	22.61 +3.09	28.24 +0.88	46.12 +1.04	29.94 +3.85	26.62 +0.92	44.69 +0.94	25.28 +3.08	**26.21** +0.90	39.79 +1.13
EvoLMM [29]	20.84 -0.70	26.71 +0.40	41.52 -0.83	18.75 -0.77	26.72 -0.64	45.86 +0.78	25.15 -0.94	26.15 +0.45	42.91 -0.84	23.04 +0.84	24.68 -0.63	39.39 +0.73
iReasoner [28]	20.93 -0.61	26.75 +0.44	41.59 -0.76	18.81 -0.71	26.75 -0.61	45.96 +0.88	25.23 -0.86	26.24 +0.54	43.00 -0.75	23.14 +0.94	24.79 -0.52	39.47 +0.81
Ours	**38.39** +16.85	**27.42** +1.11	**45.86** +3.51	**34.25** +14.73	**28.60** +1.24	**48.69** +3.61	**42.64** +16.55	**26.77** +1.07	**47.55** +3.80	**41.86** +19.66	26.20 +0.89	**41.81** +3.15
Qwen3-VL-4B-Instruct												
Base	27.35	26.20	42.95	22.36	28.91	45.95	31.10	25.64	44.20	34.54	25.26	39.70
VisPlay [6]	28.59 +1.24	26.41 +0.21	44.10 +1.15	25.46 +3.10	29.10 +0.19	45.72 -0.23	34.38 +3.28	25.52 -0.12	44.05 -0.15	31.34 -3.20	25.03 -0.23	39.42 -0.28
VisionZero-CLEVR [30]	24.69 -2.66	26.15 -0.05	42.92 -0.03	27.05 +4.69	28.71 -0.20	45.74 -0.21	36.29 +5.19	25.35 -0.29	44.36 +0.16	38.08 +3.54	25.54 +0.28	39.91 +0.21
VisionZero-Chart [30]	29.11 +1.76	26.21 +0.01	43.22 +0.27	28.12 +5.76	29.22 +0.31	46.13 +0.18	35.38 +4.28	25.81 +0.17	44.50 +0.30	32.95 -1.59	25.47 +0.21	39.91 +0.21
VisionZero-RW [30]	31.13 +3.78	26.42 +0.22	44.70 +1.75	28.61 +6.25	29.15 +0.24	46.09 +0.14	35.67 +4.57	26.03 +0.39	44.46 +0.26	**38.89** +4.35	25.45 +0.19	39.89 +0.19
EvoLMM [29]	30.53 +3.18	26.12 -0.08	42.87 -0.08	25.36 +3.00	28.61 -0.30	45.73 -0.22	28.16 -2.94	25.83 +0.19	44.25 +0.05	38.29 +3.75	25.34 +0.08	39.82 +0.12
iReasoner [28]	30.68 +3.33	26.25 +0.05	42.93 -0.02	25.52 +3.16	28.73 -0.18	45.78 -0.17	28.24 -2.86	25.93 +0.29	44.36 +0.16	38.41 +3.87	25.47 +0.21	39.97 +0.27
Ours	**39.65** +12.30	**29.52** +3.32	**46.91** +3.96	**34.97** +12.61	**29.60** +0.69	**48.86** +2.91	**37.17** +6.07	**26.64** +1.00	**47.82** +3.62	38.59 +4.05	**26.43** +1.17	**42.05** +2.35
Qwen3-VL-8B-Instruct												
Base	29.01	27.25	45.72	24.46	29.91	47.82	34.02	26.69	46.55	36.21	26.28	41.62
VisPlay [6]	31.02 +2.01	27.58 +0.33	45.69 -0.03	26.41 +1.95	29.98 +0.07	47.96 +0.14	35.01 +0.99	26.77 +0.08	46.78 +0.23	36.73 +0.52	26.24 -0.04	41.89 +0.27
VisionZero-CLEVR [30]	34.88 +5.87	28.36 +1.11	46.08 +0.36	29.92 +5.46	30.14 +0.23	48.44 +0.62	35.73 +1.71	26.83 +0.14	46.93 +0.38	37.21 +1.00	26.53 +0.25	41.97 +0.35
VisionZero-Chart [30]	36.12 +7.11	28.62 +1.37	46.21 +0.49	31.47 +7.01	30.22 +0.31	48.53 +0.71	33.88 -0.14	26.75 +0.06	46.82 +0.27	35.98 -0.23	26.39 +0.11	41.88 +0.26
VisionZero-RW [30]	37.42 +8.41	28.71 +1.46	46.24 +0.52	33.87 +9.41	30.27 +0.36	48.59 +0.77	37.92 +3.90	27.03 +0.34	47.39 +0.84	38.03 +1.82	26.96 +0.68	42.63 +1.01
EvoLMM [29]	29.84 +0.83	27.21 -0.04	45.78 +0.06	24.89 +0.43	29.95 +0.04	48.31 +0.49	33.74 -0.28	26.65 -0.04	47.02 +0.47	36.05 -0.16	26.31 +0.03	41.73 +0.11
iReasoner [28]	33.26 +4.25	28.04 +0.79	45.67 -0.05	28.44 +3.98	30.09 +0.18	48.21 +0.39	36.34 +2.32	26.61 -0.08	47.06 +0.51	37.48 +1.27	26.63 +0.35	42.21 +0.59
Ours	**38.49** +9.48	**28.93** +1.68	**46.31** +0.59	**34.98** +10.52	**30.31** +0.40	**48.66** +0.84	**38.62** +4.60	**27.11** +0.42	**47.61** +1.06	**38.42** +2.21	**27.11** +0.83	**42.78** +1.16
Qwen3-VL-32B-Instruct												
Base	33.45	29.72	49.85	27.74	33.38	50.02	38.68	29.73	49.83	41.25	29.68	45.85
VisPlay [6]	35.01 +1.56	30.01 +0.29	49.79 -0.06	28.62 +0.88	33.51 +0.13	50.09 +0.07	39.02 +0.34	29.81 +0.08	49.96 +0.13	41.63 +0.38	29.61 -0.07	45.96 +0.11
VisionZero-CLEVR [30]	38.74 +5.29	30.62 +0.90	50.94 +1.09	32.41 +4.67	34.02 +0.64	50.61 +0.59	39.64 +0.96	30.31 +0.58	51.02 +1.19	42.31 +1.06	29.92 +0.24	46.21 +0.36
VisionZero-Chart [30]	39.82 +6.37	30.81 +1.09	50.88 +1.03	33.14 +5.40	34.10 +0.72	50.66 +0.64	38.52 -0.16	30.44 +0.71	50.74 +0.91	41.11 -0.14	29.88 +0.20	46.12 +0.27
VisionZero-RW [30]	41.21 +7.76	30.94 +1.22	51.08 +1.23	35.12 +7.38	34.29 +0.91	50.69 +0.67	40.05 +1.37	30.76 +1.03	51.71 +1.88	42.84 +1.59	30.27 +0.59	46.41 +0.56
EvoLMM [29]	34.01 +0.56	29.63 -0.09	49.92 +0.07	28.03 +0.29	33.44 +0.06	50.28 +0.26	38.44 -0.24	29.69 -0.04	50.31 +0.48	41.02 -0.23	29.71 +0.03	45.93 +0.08
iReasoner [28]	37.62 +4.17	30.48 +0.76	50.61 +0.76	31.68 +3.94	33.97 +0.59	50.53 +0.51	39.73 +1.05	30.18 +0.45	50.96 +1.13	42.19 +0.94	29.94 +0.26	46.19 +0.34
Ours	**42.17** +8.72	**31.02** +1.30	**51.23** +1.38	**35.95** +8.21	**34.41** +1.03	**50.74** +0.72	**40.32** +1.64	**30.89** +1.16	**52.02** +2.19	**43.06** +1.81	**30.34** +0.66	**46.54** +0.69

Training Data. We use 4000 raw, unlabeled images sampled from COCO train2014/train2017 [13], with no captions, bounding boxes, or semantic labels retained, and evaluate on disjoint validation splits with zero image overlap. Spatial transformations (affine, crop, and flip) are applied online during training to generate geometric invariance targets. We also report Objects365 [25] training results showing consistent gains beyond COCO-specific image exposure.

Baselines and Evaluation. We compare against five self-evolving baselines on the same base models: VisPlay [6], EvoLMM [29], and iReasoner [28], which are fully unsupervised and require no external rewards or annotations, and VisionZero [30] (CLEVR, Chart, and Real-World [RW] variants), where only CLEVR is label-free, while Chart and Real-World use GPT-4o during dataset construction. All evaluations are run on AMD MI250X GPUs using the lmms-eval framework [35], with HuggingFace Transformers v4.38 and bfloat16 precision for consistency with training. We evaluate on four image captioning benchmarks (COCO [2014/2017 average] [13], NoCaps [1], Flickr30k [22], TextCaps [26]), twelve VQA and reasoning benchmarks (GQA [8], OK-VQA [19], VQAv2 [5], AI2D [9], ChartQA [20], InfoVQA [21], ScienceQA [18], MMMU [34], Cap-

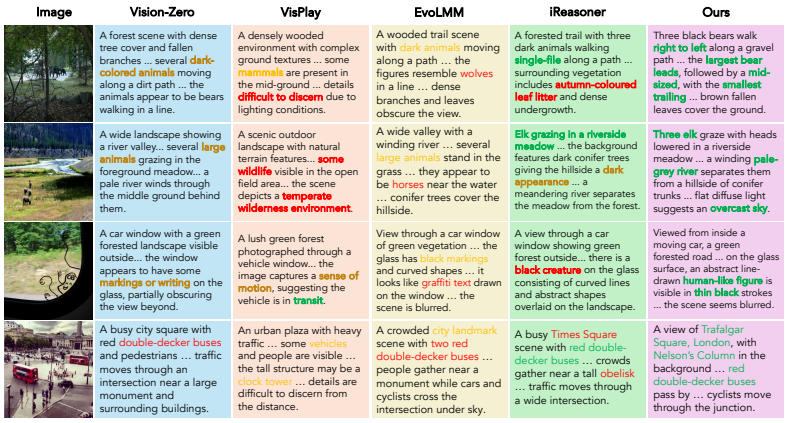

Fig. 6: Qualitative comparison of VISE and baselines on four images. Baselines generate either vague category-level descriptions ("large animals," "vehicles") or confident hallucinations ("wolves," "obelisk"), reflecting reliance on language priors. In contrast, VISE provides specific, image-grounded descriptions: it identifies the three bears by size and position, names the elk and river color, recognizes the human-like figure on the car window, and correctly identifies Trafalgar Square and Nelson's Column. These differences are supported by CIDEr gains in Table 1, indicating that invariance-based training encourages reliance on visible evidence rather than scene-level priors.

tionQA [33], RWQA [32], ESB [4], MMBench [16]), and two hallucination benchmarks (POPE [12] and COCO Cap Chair [24]).

Image Captioning. In Table 1, we compare VISE with all baselines across four captioning benchmarks and four Qwen3-VL scales. Consistency-based methods trained on math or scientific images show a clear 2B pattern: EvoLMM drops by −0.70 CIDEr on COCO, −0.77 on NoCaps, and −0.94 on Flickr30k, with iReasoner showing a similar trend. This persists across scales and cannot be explained by distribution mismatch alone: when answer agreement is the reward, the model is not required to describe what it sees, so captioning suffers. VisionZero-RW is the strongest 2B captioning baseline (+4.04 COCO CIDEr) due to closer natural-scene alignment, while its Chart/CLEVR variants give smaller, less consistent gains. VisPlay improves COCO by +2.31 but regresses on NoCaps (−0.38) and TextCaps (−0.09), suggesting diversity/difficulty rewards favor question complexity over visual faithfulness. At larger scales, VisionZero variants recover further (e.g., RW reaches +8.41 COCO CIDEr at 8B), whereas EvoLMM and iReasoner remain inconsistent, pointing to distribution proximity rather than visual conditioning.

VISE improves Qwen3-VL-2B CIDEr by +16.85 on COCO, +14.73 on NoCaps, +16.55 on Flickr30k, and +19.66 on TextCaps, with gains 4×–7× larger than the strongest baseline on each benchmark and no regressions on any dataset or scale. Gains shrink with scale (+16.85 at 2B to +8.72 at 32B on COCO), con-

Table 2: Evaluation results on twelve VQA and reasoning benchmarks. All values are Accuracy except CaptionQA (GPT Score). Domain-specific baselines exhibit a consistent generalization tradeoff: EvoLMM and iReasoner improve on ScienceQA (+3.59, +3.70) but drop on OK-VQA (−2.73, −2.63), while VisionZero variants show the reverse pattern depending on training domain. VISE improves all twelve benchmarks simultaneously at 2B (+4.19 ScienceQA, +2.41 InfoVQA, +1.75 MMMU) with no regressions at any scale, demonstrating that strengthening visual grounding generalizes across task formats without domain-specific adaptation.

Method	GQA	OK-VQA	VQAv2	AI2D	ChartQA	InfoVQA	ScienceQA	MMMU	CaptionQA	RWQA	ESB	MMBench
Qwen3-VL-2B-Instruct												
Base	58.25	40.76	78.37	73.67	79.16	69.02	79.42	38.92	77.04	63.41	68.54	74.48
VisPlay [6]	58.65+0.40	41.15+0.39	78.59+0.22	74.14+0.47	79.46+0.30	69.96+0.94	80.74+1.32	39.27+0.35	77.35+0.31	63.62+0.21	68.56+0.02	74.52+0.04
VisionZero-CLEVR [30]	58.98+0.73	**41.39**+0.63	**79.04**+0.67	75.32+1.65	79.78+0.62	70.93+1.91	81.96+2.54	39.58+0.66	77.69+0.65	63.75+0.34	69.72+1.18	75.07+0.59
VisionZero-Chart [30]	58.64+0.39	40.31-0.45	78.85+0.48	74.10+0.43	**80.12**+0.96	69.61+0.59	79.93+0.51	39.50+0.58	76.54-0.50	63.64+0.23	68.52-0.02	74.51+0.03
VisionZero-RW [30]	57.91-0.34	40.17-0.59	78.89+0.52	74.20+0.53	78.69-0.47	69.58+0.56	79.90+0.48	39.27+0.35	76.69-0.35	64.31+0.90	69.77+1.23	74.40-0.08
EvoLMM [29]	59.01+0.76	38.03-2.73	77.86-0.51	75.78+2.11	79.80+0.64	70.69+1.67	83.01+3.59	39.08+0.16	76.73-0.31	63.78+0.37	69.32+0.78	74.62+0.14
iReasoner [28]	59.13+0.88	38.13-2.63	77.94-0.43	75.97+2.30	79.96+0.80	70.82+1.80	83.12+3.70	39.11+0.19	76.82-0.22	63.94+0.53	69.67+1.13	74.75+0.27
Ours	**59.41**+1.16	41.24+0.48	78.54+0.17	**76.42**+2.75	80.08+0.92	**71.43**+2.41	**83.61**+4.19	**40.67**+1.75	**79.16**+2.12	**64.58**+1.17	**70.14**+1.60	**76.72**+2.24
Qwen3-VL-4B-Instruct												
Base	60.32	47.86	80.07	80.10	83.18	77.73	87.51	45.17	82.93	67.82	76.79	83.42
VisPlay [6]	60.06-0.26	48.04+0.18	79.89-0.18	80.34+0.24	82.96-0.22	77.96+0.23	87.30-0.21	45.42+0.25	84.16+1.23	68.67+0.85	76.11-0.68	83.47+0.05
VisionZero-CLEVR [30]	61.30+0.98	48.41+0.55	81.07+1.00	80.81+0.71	83.68+0.50	78.28+0.55	88.46+0.95	46.27+1.10	83.89+0.96	68.43+0.61	77.02+0.23	84.83+1.41
VisionZero-Chart [30]	61.39+1.07	48.50+0.64	81.09+1.02	81.76+1.66	84.38+1.20	78.74+1.01	88.50+0.99	46.20+1.03	83.90+0.97	67.91+0.09	76.25-0.54	83.78+0.36
VisionZero-RW [30]	60.51+0.19	48.00+0.14	80.24+0.17	80.31+0.21	82.99-0.19	77.94+0.21	87.73+0.22	45.41+0.24	83.10+0.17	69.74+1.92	77.06+0.27	83.32-0.10
EvoLMM [29]	61.20+0.88	46.72-1.14	80.97+0.90	81.13+1.03	83.52+0.34	78.58+0.85	88.46+0.95	46.02+0.85	83.41+0.48	68.65+0.83	76.94+0.15	84.21+0.79
iReasoner [28]	61.32+1.00	46.89-0.97	81.10+1.03	81.22+1.12	83.72+0.54	78.76+1.03	88.62+1.11	46.13+0.96	83.52+0.59	68.92+1.10	76.98+0.19	84.34+0.92
Ours	**61.82**+1.50	**48.81**+0.95	**81.16**+1.09	**82.16**+2.06	**84.96**+1.78	**81.45**+3.72	**90.04**+2.53	**48.89**+3.72	**85.19**+2.26	**70.46**+2.64	**77.46**+0.67	**84.51**+1.09
Qwen3-VL-8B-Instruct												
Base	61.54	49.84	81.81	83.31	84.87	81.23	90.88	50.12	85.21	69.28	77.66	84.71
VisPlay [6]	61.66+0.12	49.93+0.09	81.78-0.03	83.45+0.14	84.79-0.08	81.36+0.13	91.02+0.14	52.61+2.49	85.32+0.11	69.41+0.13	78.24+0.58	84.62-0.09
VisionZero-CLEVR [30]	62.12+0.58	50.31+0.47	**82.72**+0.91	83.98+0.67	85.18+0.31	82.44+1.21	92.35+1.47	**52.89**+2.77	86.01+0.80	69.88+0.60	77.96+0.30	85.11+0.40
VisionZero-Chart [30]	62.05+0.51	50.26+0.42	82.69+0.88	84.04+0.73	**85.49**+0.62	82.31+1.08	92.41+1.53	49.98-0.14	85.96+0.75	69.95+0.67	76.94-0.72	85.07+0.36
VisionZero-RW [30]	61.71+0.17	49.89+0.05	81.92+0.11	83.37+0.06	84.94+0.07	81.44+0.21	91.18+0.30	50.43+0.31	85.38+0.17	69.33+0.05	78.02+0.36	84.76+0.05
EvoLMM [29]	61.94+0.40	50.18+0.34	82.38+0.57	83.89+0.58	85.05+0.18	82.02+0.79	91.96+1.08	51.56+1.44	85.82+0.61	69.74+0.46	77.51-0.15	85.02+0.31
iReasoner [28]	62.02+0.48	50.24+0.40	82.46+0.65	83.96+0.65	85.12+0.25	82.15+0.92	92.11+1.23	51.83+1.71	85.89+0.68	69.81+0.53	77.82+0.16	85.09+0.38
Ours	**62.43**+0.89	**50.61**+0.77	82.65+0.84	**84.10**+0.79	85.41+0.54	**82.83**+1.60	**92.81**+1.93	52.69+2.57	**86.35**+1.14	**70.03**+0.75	**78.62**+0.96	**85.46**+0.75
Qwen3-VL-32B-Instruct												
Base	62.08	51.16	83.56	86.63	86.00	87.77	95.68	59.47	88.96	78.16	81.12	86.65
VisPlay [6]	62.19+0.11	51.23+0.07	83.52-0.04	86.71+0.08	85.94-0.06	87.89+0.12	95.80+0.12	59.18-0.29	89.04+0.08	78.24+0.08	81.18+0.06	86.58-0.07
VisionZero-CLEVR [30]	62.71+0.63	51.63+0.47	**83.93**+0.37	87.10+0.47	87.02+1.02	88.32+0.55	96.05+0.37	60.12+0.65	89.74+0.78	79.29+1.13	81.63+0.51	87.18+0.53
VisionZero-Chart [30]	62.64+0.56	51.57+0.41	83.90+0.34	87.16+0.53	86.94+0.94	88.19+0.42	96.11+0.43	60.04+0.57	89.69+0.73	79.34+1.18	81.55+0.43	87.13+0.48
VisionZero-RW [30]	62.25+0.17	51.22+0.06	83.61+0.05	86.69+0.06	86.05+0.05	87.94+0.17	95.84+0.16	59.33-0.14	89.09+0.13	78.31+0.15	81.21+0.09	86.72+0.07
EvoLMM [29]	62.49+0.41	51.48+0.32	83.80+0.24	87.02+0.39	86.83+0.83	88.05+0.28	95.98+0.30	59.92+0.45	89.58+0.62	78.98+0.82	81.49+0.37	87.02+0.37
iReasoner [28]	62.56+0.48	51.53+0.37	83.84+0.28	87.08+0.45	86.90+0.90	88.12+0.35	96.03+0.35	59.85+0.38	89.66+0.70	79.05+0.89	81.58+0.46	87.08+0.43
Ours	**62.95**+0.87	**51.76**+0.60	83.85+0.29	**87.24**+0.61	**87.27**+1.27	**88.43**+0.66	**96.21**+0.53	**62.79**+3.32	**89.96**+1.00	**79.51**+1.35	**82.39**+1.27	**87.27**+0.62

sistent with larger models starting from stronger pretrained visual conditioning. Figure 6 illustrates the gap: baselines remain vague ('large animals near a river") or plausible but wrong ("wolves", "obelisk"), while VISE identifies three differently sized bears walking in order, a hand-drawn figure on the car window, and Trafalgar Square rather than a generic landmark. This output gap is precisely what the CIDEr gap in Table 1 reflects.

VQA and Reasoning. Table 2 reports performance across twelve VQA and reasoning benchmarks. The baselines show a consistent generalization tradeoff absent in VISE. VisionZero-Chart improves ChartQA by +0.96 at 2B but drops on OK-VQA (−0.45) and CaptionQA (−0.50); VisionZero-RW shows the reverse, gaining on RWQA (+0.90) but dropping on ChartQA (−0.47); and VisionZero-CLEVR improves structured reasoning (+2.54 ScienceQA, +1.91 InfoVQA) while remaining inconsistent on open-ended VQA. This bidirectional pattern suggests domain-specific consistency training learns training-distribution regularities alongside visual conditioning, limiting generalization unless the reward changes. EvoLMM and iReasoner follow the same trend, with strong Sci-

Table 3: Evaluation results on POPE and COCO Cap Chair hallucination benchmarks. ↓ indicates lower is better. EvoLMM and iReasoner modestly reduce Chair scores (−0.23 Chair-I at 2B) but simultaneously drop POPE accuracy (−1.42, −1.31), indicating inconsistent visual grounding. In contrast, VISE reduces Chair-I from 13.21 → 8.21 (−5.00) and Chair-S from 45.96 → 40.51 (−5.45) at 2B, surpassing the strongest baseline by +2.01 and +1.40, while also improving POPE by +1.02.

Method	POPE				COCO Cap Chair		
	Acc	F1	Prec	Recall	Chair-I↓	Chair-S↓	Cap Recall
Qwen3-VL-2B-Instruct							
Base	89.01	88.37	92.45	84.63	13.2133	45.9601	72.0939
VisPlay [6]	89.32+0.31	88.72+0.35	92.86+0.41	84.93+0.30	13.2168+0.0035	46.1869+0.2268	71.3239-0.7700
VisionZero-CLEVR [30]	89.65+0.64	89.01+0.64	93.09+0.64	85.27+0.64	11.9074-1.3059	44.9661-0.9940	72.3015+0.2076
VisionZero-Chart [30]	89.35+0.34	88.75+0.38	91.86-0.59	**85.84**+1.21	12.4823-0.7310	43.3741-2.5860	71.8795-0.2144
VisionZero-RW [30]	88.70-0.31	87.84-0.53	93.05+0.60	83.18-1.45	10.2192-2.9941	41.9136-4.0465	72.0988+0.0049
EvoLMM [29]	87.59-1.42	88.51+0.14	92.05-0.40	85.17+0.54	12.9851-0.2282	44.2238-1.7363	70.9924-1.1015
iReasoner [28]	87.70-1.31	88.56+0.19	92.12-0.33	85.26+0.63	12.9835-0.2298	44.2205-1.7396	70.9931-1.1008
Ours	**90.03**+1.02	**89.22**+0.85	**93.13**+0.68	85.63+1.00	**8.2132**-5.0001	**40.5097**-5.4504	**72.3145**+0.2206
Qwen3-VL-4B-Instruct							
Base	89.73	88.45	92.57	84.68	12.9141	44.5792	74.8582
VisPlay [6]	89.42-0.31	88.92+0.47	92.89+0.32	84.99+0.31	13.2158+0.3017	46.1849+1.6057	74.3239-0.5343
VisionZero-CLEVR [30]	89.82+0.09	88.02-0.43	93.71+1.14	82.98-1.70	12.6038-0.3103	45.3978+0.8186	74.8559-0.0023
VisionZero-Chart [30]	88.75-0.98	88.77+0.32	92.79+0.22	84.93+0.25	12.2069-0.7072	44.4025-0.1767	75.1604+0.3022
VisionZero-RW [30]	89.67-0.06	**89.25**+0.80	**93.80**+1.23	85.12+0.44	11.9079-1.0062	43.4024-1.1768	74.8604+0.0022
EvoLMM [29]	89.53-0.20	88.56+0.11	93.07+0.50	84.39-0.29	12.2031-0.7110	44.9048+0.3256	75.0184+0.1602
iReasoner [28]	89.64-0.09	88.59+0.14	93.15+0.58	84.46-0.22	12.2024-0.7117	44.9035+0.3243	75.0196+0.1614
Ours	**89.86**+0.13	88.98+0.53	92.47-0.10	**85.74**+1.06	**11.9041**-1.0100	**43.0197**-1.5595	**75.7404**+0.8822
Qwen3-VL-8B-Instruct							
Base	89.91	88.61	92.71	84.86	11.2047	43.4237	75.4236
VisPlay [6]	90.23+0.32	89.15+0.54	**92.73**+0.02	85.84+0.98	10.9364-0.2683	42.0186-1.4051	75.7562+0.3326
VisionZero-CLEVR [30]	90.18+0.27	89.06+0.45	92.72+0.01	85.68+0.81	10.9829-0.2218	42.2564-1.1673	75.7014+0.2778
VisionZero-Chart [30]	89.95+0.04	88.69+0.08	92.70-0.01	85.01+0.15	11.1628-0.0419	43.2015-0.2222	75.4728+0.0492
VisionZero-RW [30]	90.02+0.11	88.80+0.19	92.72+0.01	85.20+0.34	11.1094-0.0953	42.9624-0.4613	75.5316+0.1080
EvoLMM [29]	90.13+0.22	88.98+0.37	**92.73**+0.02	85.52+0.66	11.0187-0.1860	42.4872-0.9365	75.6441+0.2205
iReasoner [28]	90.07+0.16	88.89+0.28	92.72+0.01	85.36+0.50	11.0712-0.1335	42.7358-0.6879	75.5893+0.1657
Ours	**90.32**+0.41	**89.32**+0.71	**92.73**+0.02	**86.15**+1.29	**10.8375**-0.3672	**41.5336**-1.8901	**75.8369**+0.4133
Qwen3-VL-32B-Instruct							
Base	90.35	89.45	92.96	86.20	10.8543	42.6321	76.5346
VisPlay [6]	90.45+0.10	89.57+0.12	93.07+0.11	86.32+0.12	10.7564-0.0979	42.3891-0.2430	76.6943+0.1597
VisionZero-CLEVR [30]	90.73+0.38	89.86+0.41	93.39+0.43	86.59+0.39	10.5468-0.3075	41.9264-0.7057	77.2193+0.6847
VisionZero-Chart [30]	90.51+0.16	89.64+0.19	93.14+0.18	86.39+0.19	10.7085-0.1458	42.2678-0.3643	76.8125+0.2779
VisionZero-RW [30]	90.66+0.31	89.80+0.35	93.31+0.35	86.54+0.34	10.5994-0.2549	42.0032-0.6289	77.0816+0.5470
EvoLMM [29]	90.59+0.24	89.72+0.27	93.23+0.27	86.46+0.26	10.6521-0.2022	42.1346-0.4975	76.9348+0.4002
iReasoner [28]	90.39+0.04	89.49+0.04	92.95-0.01	86.28+0.08	10.8129-0.0414	42.5184-0.1137	76.5812+0.0466
Ours	**90.81**+0.46	**89.92**+0.47	**93.45**+0.49	**86.65**+0.45	**10.4173**-0.4370	**41.8735**-0.7586	**77.4621**+0.9275

enceQA gains (+3.59, +3.70 at 2B) but drops on OK-VQA (−2.73, −2.63) and VQAv2 (−0.51, −0.43), consistent with the captioning dips in Table 1. VISE improves all twelve benchmarks at 2B, including +4.19 ScienceQA, +2.41 InfoVQA, +1.75 MMMU, and +2.12 CaptionQA, with no regressions at any scale. At 4B, its MMMU gain reaches +3.72, the largest on that benchmark across all methods and scales, indicating that stronger conditioning especially benefits multi-discipline visual reasoning. Across four scales, VISE avoids the structured-versus-open-ended tradeoff of prior methods: rather than adapting to a domain, its invariance reward improves visual conditioning across task formats.

Hallucination. Table 3 evaluates all methods on POPE and COCO Cap Chair. VisPlay increases hallucination at 2B (Chair-I +0.003, Chair-S +0.23), consistent with diversity rewards prioritizing question complexity over visual fidelity. EvoLMM and iReasoner reduce Chair scores modestly (−0.23, −1.74) but drop POPE accuracy (−1.42, −1.31) simultaneously. Sentence-level hallucinations decline while binary object-presence reliability weakens, pointing to inconsistent rather than substantive improvement. VisionZero-RW is the strongest baseline (Chair-I −2.99, Chair-S −4.05), due to broader real-world visual coverage. VISE reduces Chair-I by −5.00 and Chair-S by −5.45 at 2B, surpassing the best baseline by +2.01 and +1.40 while also improving POPE by +1.02. This joint improvement across both metrics is unique to VISE: penalizing confident predictions under regional perturbation discourages generating objects that are statistically plausible but not visually present. Gains attenuate with model size (Chair-I −0.37, −0.44 at 8B, 32B), in-line with the captioning trend in Table 1.

Effect of Model Scale. Gains from VISE are largest at smaller scales and attenuate with model size, consistent across all task groups. We attribute this to a capacity ceiling effect in the self-evolving setting: larger models enter post-training with stronger conditioning consolidated during pretraining and instruction tuning, leaving less headroom for invariance-based correction. This aligns with prior observations that self-evolving methods exhibit diminishing returns at scale [6], and suggests that invariance rewards are most effective where visual under-conditioning remains pronounced, particularly in smaller models that have not yet developed robust evidence-binding behavior.

Backbone Generalization. Table 4 evaluates VISE on four architecturally diverse backbones trained under an identical setup using the same 4000 unlabeled COCO images. Captioning and hallucination gains are consistent across families, with NoCaps showing the largest absolute improvements. Reasoning and VQA also improve without regressions, including on Llama-3.2-11B (+1.31 POPE) despite its weaker baseline grounding, indicating that the reward is insensitive to pretraining distribution. The consistency of improvements across architectures confirms that the invariance reward is architecture-agnostic and that visual under-conditioning is a general phenomenon addressed by VISE regardless of backbone.

Domain and Adaptation Analysis. Table 5 analyzes two possible sources of training effects on Qwen3-VL-8B over three seeds: image domain and adaptation strategy. For domain, VISE is trained on either COCO [13] or Objects365 [25] images, always without captions, boxes, or labels; COCO training uses train2014/train2017 and evaluates on disjoint validation splits with zero image overlap. Objects365 training gives nearly identical gains, showing that improvements are not driven by COCO-specific image exposure: LoRA reaches 38.49/38.62 COCO/Flickr30k with COCO training and 38.57/38.71 with Objects365 training. For adaptation, the same table compares full fine-tuning (FFT) with LoRA

Table 4: Effectiveness of VISE across four diverse LMM backbones. VISE consistently improves captioning across all families, with COCO CIDEr gains of +9.48, +9.01, +7.65, and +6.44, and larger NoCaps gains of +10.52, +10.16, +8.67, and +6.25. Hallucination and VQA also improve without regressions, showing that VISE is architecture-agnostic and addresses a general visual-prior-driven generation failure.

Model	Captioning				Hallucination			VQA		Reasoning			
	COCO	NoCaps	Flickr30k	TextCaps	Chair-I↓	Chair-S↓	$POPE_{F1}$	GQA	CaptionQA	AI2D	ChartQA	InfoVQA	ScienceQA
Qwen3-VL-8B-Instruct													
Base	29.01	24.46	34.02	36.21	11.20	43.42	88.61	61.54	85.21	83.31	84.87	81.23	90.88
Ours	**38.49**+9.48	**34.98**+10.52	**38.62**+4.60	**38.42**+2.21	**10.84**-0.36	**41.53**-1.89	**89.32**+0.71	**62.43**+0.89	**86.35**+1.14	**84.10**+0.79	**85.41**+0.54	**82.83**+1.60	**92.81**+1.93
InternVL3-8B-Instruct													
Base	28.43	23.81	33.47	35.62	11.43	43.87	90.83	61.12	84.73	83.19	82.40	68.77	97.19
Ours	**37.44**+9.01	**33.97**+10.16	**37.53**+4.06	**37.94**+2.32	**11.14**-0.29	**42.31**-1.56	**91.64**+0.81	**61.94**+0.82	**85.96**+1.23	**83.61**+0.42	**82.76**+0.36	**69.31**+0.54	**97.77**+0.58
Gemma3-12B-It													
Base	22.18	18.94	27.53	24.06	13.57	46.83	80.65	56.38	74.92	79.05	55.64	50.69	83.41
Ours	**29.83**+7.65	**27.61**+8.67	**31.47**+3.94	**25.93**+1.87	**13.28**-0.29	**45.37**-1.46	**81.69**+1.04	**57.14**+0.76	**75.88**+0.96	**79.38**+0.33	**55.97**+0.33	**50.94**+0.25	**83.89**+0.48
Llama-3.2-11B-Vision-Instruct													
Base	16.37	14.22	21.84	18.53	16.24	52.17	75.83	51.63	67.44	46.44	29.24	56.69	56.87
Ours	**22.81**+6.44	**20.47**+6.25	**25.63**+3.79	**20.11**+1.58	**15.97**-0.27	**51.03**-1.14	**77.14**+1.31	**52.29**+0.66	**68.31**+0.87	**46.71**+0.27	**29.48**+0.24	**56.93**+0.24	**57.43**+0.56

Table 5: Training-domain and tuning-strategy validation on Qwen3-VL-8B-Instruct (mean±std over 3 seeds). VISE obtains consistent gains when trained on either COCO or Objects365 images. LoRA also outperforms full fine-tuning (FFT), supporting our frozen-encoder training design.

Category	Training setup	Captioning		Hallucination	VQA	Reasoning
		COCO	Flickr30k	$POPE_{acc}$	CaptionQA	ScienceQA
Baseline	Qwen3-VL-8B-Instruct	29.01±0.71	34.02±0.48	89.91±0.58	85.21±0.29	90.88±0.61
FFT	COCO Training	32.80±0.45	35.86±0.61	90.07±0.43	85.47±0.58	91.45±0.82
	Objects365 Training	33.51±0.50	35.74±0.51	90.19±0.46	85.64±0.44	91.48±0.55
LoRA	COCO Training	**38.49±0.67**	**38.62±0.52**	90.32±0.45	86.35±0.28	92.81±0.66
	Objects365 Training	**38.57±0.36**	**38.71±0.46**	**90.34±0.68**	**86.52±0.67**	**92.93±0.28**

under both domains. FFT improves over the base model but consistently underperforms LoRA across captioning, hallucination, VQA, and reasoning, supporting our design choice to avoid noisy unsupervised encoder updates and instead adapt the cross-modal and decoder components.

Ablation Study. Table 6 isolates the contribution of each reward component on Qwen3-VL-2B and Qwen3-VL-8B.

Training with R_{geo} yields moderate captioning gains (+4.83 COCO CIDEr, +4.33 Flickr30k at 2B) and modest hallucination reductions (Chair-I −1.35, Chair-S −1.45), highlighting that spatial consistency provides a meaningful visual signal. However, because it cannot penalize visually unsupported but plausible generations, its improvements remain well below those of the full model. R_{sem} accounts for most of VISE's gains, delivering +13.99 COCO CIDEr and reducing Chair-I by −4.15 and Chair-S by −4.45 at 2B. This is consistent with our findings that hallucination reduction stems directly from the semantic reward design. By penalizing confident predictions under regional perturbation, the ghosting signal emerges as the primary driver of the captioning and hallucination gains in Tables 1 and 3.

Table 6: Ablation of VISE reward components. R_{geo} and R_{sem} denote geometric- and semantic-invariance rewards; COCO CIDEr is averaged over the 2014/2017 val splits. At 2B, R_{geo} gives moderate gains (+4.83 CIDEr, Chair-I −1.35), while R_{sem} drives most improvements (+13.99 CIDEr, Chair-I −4.15). Combining both further improves over R_{sem} (+2.86 CIDEr, Chair-I −0.85), confirming complementary spatial-consistency and evidence-sensitivity signals.

Method	Captioning				Hallucination			VQA		Reasoning			
	COCO	NoCaps	Flickr30k	TextCaps	Chair-I↓	Chair-S↓	POPE	GQA	CaptionQA	AI2D	ChartQA	InfoVQA	ScienceQA
Qwen3-VL-2B-Instruct													
Base	21.54	19.52	26.09	22.20	13.21	45.96	89.01	58.25	77.04	73.67	79.16	69.02	79.42
R_{geo} only	26.37+4.83	23.07+3.55	30.42+4.33	27.42+5.22	11.86-1.35	44.51-1.45	89.29+0.28	58.56+0.31	77.61+0.57	74.21+0.54	79.53+0.37	69.71+0.69	80.18+0.76
R_{sem} only	35.53+13.99	31.75+12.23	39.83+13.74	38.52+16.32	9.06-4.15	41.51-4.45	89.86+0.85	59.21+0.96	78.80+1.76	75.08+1.41	79.82+0.66	70.61+1.59	81.94+2.52
Full (Ours)	38.39+16.85	34.25+14.73	42.64+16.55	41.86+19.66	8.21-5.00	40.51-5.45	90.03+1.02	59.41+1.16	79.16+2.12	76.42+2.75	80.08+0.92	71.43+2.41	83.61+4.19
Qwen3-VL-8B-Instruct													
Base	29.01	24.46	34.02	36.21	11.20	43.42	89.91	61.54	85.21	83.31	84.87	81.23	90.88
R_{geo} only	31.84+2.83	27.12+2.66	36.18+2.16	37.93+1.72	11.02-0.18	42.88-0.54	89.99+0.08	61.79+0.25	85.52+0.31	83.82+0.51	85.21+0.34	83.02+1.79	92.83+1.95
R_{sem} only	35.27+6.26	31.43+6.97	37.41+3.39	37.18+0.97	10.91-0.29	42.16-1.26	90.11+0.20	62.08+0.54	85.94+0.73	83.96+0.65	85.33+0.46	83.11+1.88	92.85+1.97
Full (Ours)	38.49+9.48	34.98+10.52	38.62+4.60	38.42+2.21	10.84-0.36	41.53-1.89	90.32+0.41	62.43+0.89	86.35+1.14	84.10+0.79	85.41+0.54	82.83+1.60	92.81+1.93

Combining both rewards yields additional and consistent gains, with R_{geo} contributing complementary benefits to R_{sem}. At 2B, the full model adds +2.86 COCO CIDEr and a further −0.85 Chair-I reduction beyond R_{sem} alone. The same pattern holds at 8B: while R_{sem} dominates captioning (+6.26 vs. +2.83 COCO CIDEr), the combined model delivers broader improvements across VQA and reasoning tasks. Together, these results show that spatial consistency and evidence sensitivity address distinct facets of visual under-conditioning and are both necessary for the full benefits of VISE.

5 Conclusion

We introduced VISE, a fully unsupervised self-evolving framework that addresses visual under-conditioning in LMMs without annotations, external reward models, or multi-role formulations. Using a single model, VISE combines geometric consistency under spatial transformations with semantic sensitivity under regional perturbation, encouraging the decoder to attend to visual tokens rather than language priors. Experiments show consistent gains across captioning, VQA, reasoning, and hallucination benchmarks with no task tradeoffs, across four model scales and four diverse backbones. Ablations show that semantic invariance drives most gains, while geometric invariance adds complementary improvements, together targeting distinct facets of the failure mode. Overall, our results show that answer-consistency rewards are insufficient for genuine visual improvement, and that directly increasing attention to visual tokens during decoding yields broad, robust gains without supervised signals.

6 Acknowledgement

The computations were enabled by resources provided by LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
3. Chen, X., Hu, W., Li, H., Zhou, J., Chen, Z., Cao, M., Zeng, Y., Zhang, K., Yuan, Y.J., Han, J., et al.: C2-evo: Co-evolving multimodal data and model for self-improving reasoning. arXiv preprint arXiv:2507.16518 (2025)
4. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 346–355 (2024)
5. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
6. He, Y., Huang, C., Li, Z., Huang, J., Yang, Y.: Visplay: Self-evolving vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26274–26284 (2026)
7. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
8. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
9. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
10. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMLR (2019)
11. Li, T., Sun, Z., Wei, J., He, C., Wu, L., Tan, C.: Decouple to generalize: Context-first self-evolving learning for data-scarce vision-language reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29357–29366 (2026)
12. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

14. Liu, J., Xiong, K., Xia, P., Zhou, Y., Ji, H., Feng, L., Han, S., Ding, M., Yao, H.: Agent0-vl: Exploring self-evolving agent for tool-integrated vision-language reasoning. arXiv preprint arXiv:2511.19900 (2025)
15. Liu, W., Li, J., Zhang, X., Zhou, F., Cheng, Y., He, J.: Diving into self-evolving training for multimodal reasoning. arXiv preprint arXiv:2412.17451 (2024)
16. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in neural information processing systems **35**, 2507–2521 (2022)
19. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
20. Masry, A., Tan, J.Q., Joty, S., Hoque, E., et al.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
21. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
22. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
23. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
24. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1437, https://aclanthology.org/D18-1437/
25. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8429–8438 (2019). https://doi.org/10.1109/ICCV.2019.00852
26. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European conference on computer vision. pp. 742–758. Springer (2020)
27. Singh, A., Co-Reyes, J.D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P.J., Harrison, J., Lee, J., Xu, K., et al.: Beyond human data: Scaling self-training for problem-solving with language models. arXiv preprint arXiv:2312.06585 (2023)
28. Sunil, M., Venmathimaran, M., Kavitha, M.S.: ireasoner: Trajectory-aware intrinsic reasoning supervision for self-evolving large multimodal models. arXiv preprint arXiv:2601.05877 (2026)

29. Thawakar, O., Venkatraman, S., Thawkar, R., Shaker, A., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.: Evolmm: Self-evolving large multimodal models with continuous rewards. arXiv preprint arXiv:2511.16672 (2025)
30. Wang, Q., Liu, B., Zhou, T., Shi, J., Lin, Y., Chen, Y., Li, H.H., Wan, K., Zhao, W.: Vision-zero: Scalable VLM self-evolution via multi-agent self-play. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=s00SNXREV6
31. Wang, X., Chen, J., Wang, Z., Zhou, Y., Zhou, Y., Yao, H., Zhou, T., Goldstein, T., Bhatia, P., Kass-Hout, T., et al.: Enhancing visual-language modality alignment in large vision language models via self-improvement. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 268–282 (2025)
32. xAI, visheratin: Realworldqa. https://huggingface.co/datasets/visheratin/realworldqa (2024), https://huggingface.co/datasets/visheratin/realworldqa
33. Yang, S., Liu, Y., Zhai, B., Sun, X., Liu, Z., Barsoum, E., Li, M., Xu, C.: Captionqa: Is your caption as useful as the image itself? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23741–23750 (2026)
34. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
35. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)